

P16234m

GENOMIC SEQUENCE OF NGR234 SYMBIOTIC
PLASMID, ITS GENE MAP, AND ITS USE IN
DIAGNOSTICS AND GENE TRANSFER IN AGRICULTURE

5

TECHNICAL FIELD

This invention relates to a symbiotic plasmid of the broad host-range *Rhizobium* sp. NGR234 and its use. In particular, this invention relates to the isolation and analysis of the complete sequence of the NGR234 symbiotic plasmid pNGR234a, and the open reading frames (ORFs) identifiable therein as well as the proteins expressible from said ORFs.

15

BACKGROUND OF THE INVENTION

Together with carbon, hydrogen and oxygen, nitrogen is one of the essential components in organic chemistry. Although it is present in vast quantities in the atmosphere, nitrogen in its diatomic form N_2 remains unassimilable by living organisms. The nitrogen cycle begins by the fixation of nitrogen into ammonia which is chemically more reactive and can be assimilated into the food chain. A large fraction of the total nitrogen fixed every year is produced by microorganisms. Among these, the soil bacteria of the genera *Azorhizobium*, *Bradyrhizobium*, *Sinorhizobium* and *Rhizobium*, generally referred to as rhizobia, fix nitrogen in symbiotic associations with many plants from the Leguminosae family. This highly specific interaction leads to the formation of specialized root-, and in the case of *Azorhizobium*, stem- structures called nodules. It is within these nodules that rhizobia differentiate into bacteroids capable of fixing atmospheric nitrogen into ammonia. In turn, ammonia diffuses into the vegetal cells and sustains plant growth even under limiting nitrogen conditions.

0993964-082704

The *Rhizobium*-legume interaction presents many interesting features. Obviously, the possibility of using this symbiosis as an "environmentally friendly" way to provide some of the most important world crops (such as soybean, bean and many other legumes) with fixed nitrogen without using nitrate-rich fertilizers, has important economic consequences. It is also an ideal model to study a non-pathogenic interaction between bacteria and a highly developed, multicellular organism such as the host plant. Furthermore, the various steps involved in the establishment of a functional nitrogen symbiosis, which include some dramatic morphological changes as well as processes of cellular differentiation, require a complex exchange of molecular signals. Despite many decades of studies, it is only recently that the *Rhizobium*-legume interaction has been partially understood at the molecular level. The establishment of a functional symbiosis can be divided into two major steps as follows.

20 (A) Rhizosphere ecology and nodulation:

Rhizobia are soil bacteria that proliferate in the rhizosphere of compatible plants, taking advantage of the many compounds released by plant roots. In return it has been shown that the presence of rhizobia in the rhizosphere reduces susceptibility of plants to many root diseases. In the case of low nitrogen levels in the soil, compatible rhizobia can interact with host plants and start the nodulation process (Long, 1989; Fellay et al., 1995; van Rhijn and Vanderleyden, 1995). Molecular signalling between the two partners begins with the release by the plant of phenolic compounds (mostly flavonoids) that induce the expression of nodulation genes (referred to as *nod*, *nol* and *noe* genes). The NodD1 gene product appears to be the central mediator between the plant signal and nodulation gene induction (Bender et al., 1988). It is modified by the binding of flavonoids and acts as a positive regulator on the expression of the remaining nodulation genes. Among

T02280-49662660

them, the *nodABC* loci encode products responsible for the synthesis of the core structure of lipooligosaccharides called Nod factors (Relić et al., 1994). More nodulation genes are involved in strain-specific modifications of the Nod factors as well as in its secretion. It seems established now that variability in the structure of Nod factors may play a significant role in the determination of the host-range of a given *Rhizobium* strain, that is in its ability to efficiently nodulate different legumes. For example, the strain *Rhizobium meliloti* can only nodulate *Medicago*, *Melilotus* and *Trigonella* ssp., whereas *Rhizobium* sp. NGR234 can symbiotically interact with more than 105 different genera of plants, including the non-legume *Parasponia andersonii*.

The structure of many Nod factors, their isolation from *Rhizobium* strains and their commercial application in agriculture have been described (NodNGR-Faktoren: Relić et al., 1994; WO 94/00466; NodRm-Faktoren: WO 91/15496). Secreted Nod factors act in turn as signal molecules that allow rhizobia to enter young root hairs of a host plant, and induce root-cortical cell division that will produce the future nodule. Invaginated rhizobia progress towards the forming nodule within infection threads that are synthesized by the plant cells. Bacteria are then released into the cytoplasm of dividing nodule cells where they differentiate into bacteroids capable of fixing atmospheric nitrogen.

With respect to regulation of the nodulation genes, other regulatory genes with similarities to *nodD1* (genes that belong to the *lysR* family) have been identified in various strains (Davis and Johnston, 1990). The function of these genes, called *nodD2*, *nodD3* or *syrM*, is only partially understood. Some *nodD* genes have been described (WO 94/00466; CA 1314249; WO 87/07910; US 5023180). Also, recombinant DNA molecules including the consensus sequence of the promoters of *nodD1*-regulated genes, called *nod-boxes* (Fisher and Long, 1993), have been disclosed (US 5484718;

T02280-4956550

US 5085588). Finally, recombinant plasmids with the *nodABC* genes or, in one case (*Bradyrhizobium japonicum*), a sequence influencing host specificity have been disclosed (US 5045461; US 4966847).

5

(B) Symbiotic nitrogen fixation:

Inside the nodules, rhizobia differentiate into bacteroids that express the enzymatic complex (nitrogenase) required for the reduction of atmospheric nitrogen into ammonia. The nitrogenase is encoded by three genes *nifH*, *nifD* and *nifK* which are well conserved in nitrogen fixing organisms (Badenoch-Jones et al., 1989). Many additional loci are necessary for functional nitrogenase activity. Those originally identified in *Klebsiella pneumoniae* are known as *nif* genes, whereas those found only in *Rhizobium* strains are described as *fix* genes (Fischer, 1994). Some of these gene products are required for the biosynthesis of cofactors, the assembly of the enzymatic complex or play regulatory and different accessory roles (oxygen-limited respiration, etc.). Many of these genes are less conserved among the various rhizobial strains and in some cases their function is still not fully understood. The high sensitivity of the nitrogenase complex to free oxygen requires a very strict control of most *nif* and *fix* gene expression. In this respect, the FixL, FixJ, FixK, NifA and RpoN proteins have been identified in representative *Rhizobium* species as the major regulatory elements that, in microanaerobic conditions, activate the synthesis of the nitrogenase complex (Fischer, 1994). Recombinant DNA molecules containing *nif* genes/promoters have been disclosed: *nifH* promoters of *B. japonicum* (US 5008194), *nifH* and *nifD* promoter of *R. japonicum* (EP 164245), *nifA* of *B. japonicum* and *R. meliloti* (EP 339830), *nifHDK* and hydrogen-uptake (*hup*) genes of *R. japonicum* (EP 205071).

Many more genetic determinants play a significant role in the *Rhizobium*-legume symbiosis. Genes (*exo*, *lps* and *ndv*

T02280-4966847

genes) involved in the production of extracellular polysaccharides (EPS), lipopolysaccharides (LPS) and cyclic glucanes of rhizobia play an essential role in the symbiotic interaction (Long et al., 1988; Stanfield et al., 1988).

5 Mutation in these genes negatively influences the development of functional nodules. In this respect, some exopolysaccharides of the NGR234 derivative strain ANU280, have been disclosed (WO 87/06796). Although Nod factors seem to play a key role in the nodulation process,

10 experimental data indicate that other signal molecules produced by the bacterial symbionts are required for functional symbiosis and may play a role in coordinating various steps such as the controlled invasion process, the release of rhizobia from the infection thread into the plant

15 cell cytoplasm, the bacteroid differentiation process, etc. Moreover, the need for rhizobia to survive in the rhizosphere and to compete adequately with other microorganisms requires many more unidentified genes that, although they may not be characterised as proper symbiotic

20 loci, do affect the efficiency of the various strains to induce functional nitrogen fixing symbiosis in field conditions. Finally, in our view genetic engineering of improved rhizobial strains cannot be pursued without a more extended knowledge of the structure and complexity of the

25 *Rhizobium* symbiotic genome.

In this respect we decided to determine the complete DNA sequence of a symbiotic plasmid of *Rhizobium* sp. NGR234. In contrast to *Bradyrhizobium* and *Azorhizobium* that carry

30 symbiotic genes on large chromosomes (ca. 8 Mbp) and to *R. meliloti* that harbours two very large symbiotic plasmids of 1.4 and 1.6 Mbp, NGR234 carries a single plasmid of ca. 500 kbp, pNGR234a. Moreover, it has been shown by transfer of pNGR234a into heterologous rhizobia, and even into non-

35 nodulating *Agrobacterium tumefaciens*, that most nodulation functions are encoded by this plasmid (Broughton et al., 1984). The fact that NGR234 is able to interact symbiotically with more plants than any other known strain,

00339964 082701

and that a complete ordered cosmid library of pNGR234a was available, reinforced NGR234 as the best choice for a large-scale sequencing effort on a symbiotic plasmid (Perret et al., 1991; Freiberg et al., 1997).

5

Automated fluorescent methods have been used to sequence cosmids from eukaryotic organisms, including *Saccharomyces cerevisiae* (Levy, 1994), *Caenorhabditis elegans* (Sulston et al., 1992), *Drosophila melanogaster* (Hartl and Palazzolo, 1993), and *Homo sapiens* (Bodmer, 1994), as well as chromosomes from the prokaryotes *Haemophilus influenzae* (Fleischmann et al., 1995) and *Mycoplasma genitalium* (Fraser et al., 1995). In most large-scale sequencing centres this technology is based mainly on the shotgun approach. After random fragmentation of DNA (e.g. cosmids, bacterial artificial chromosomes (BACs), entire chromosomes) using sonication or mechanical forces, size-selected fragments are subcloned into M13 phages, phagemids or plasmids and sequenced by cycle sequencing using dye primers (Craxton, 1993). A disadvantage of this method is that DNA regions with elevated GC contents produce large numbers of compressions (unresolvable foci in sequence gels) in the dye primer sequences leading to several hundred compressions per assembled cosmid sequence. It is known that the use of dye terminators - fluorescently labelled dideoxynucleoside triphosphates - instead of dye primers reduces the number of compressions (Rosenthal and Charnock-Jones, 1993). Therefore, dye terminators are frequently being used for gap closure and proofreading after assembly of the shotgun data.

To sequence GC-rich cosmids with the highest accuracy, the effectiveness of shotgun sequencing with dye terminators in comparison to dye primer sequencing was investigated. To improve the incorporation of dye terminators into DNA, a modified *Taq* DNA polymerase carrying a single mutation was used (Tabor and Richardson, 1995). This enzyme has properties similar to a thermostable "sequenase" and is

commercially available as Thermo Sequenase (Amersham, Buckinghamshire, UK) or AmpliTaq FS (Perkin-Elmer, Foster City, CA, USA). Concentrations of dye terminators needed in the cycle sequencing reactions can be reduced by 20 - 250 times. It was found that dye terminator shotgun sequencing leads to compression-free raw data that can be assembled much faster than shotgun data mainly obtained by dye primer sequencing. This strategy thus allows a several-fold increase in speed to sequence individual cosmids. This was demonstrated by comparing assembly of the sequence data of two cosmids from pNGR234a generated by different chemistries: Cosmid pXB296 was sequenced with dye terminators, whereas data for pXB110 were obtained using the common dye primer method. Also disclosed is the analysis of the entire pXB296 sequence.

Moreover, the dye terminator shotgun sequencing strategy used to generate the sequence data for pXB296 was also used to sequence all the other remaining overlapping cosmids of the plasmid pNGR234a. In summary, 20 cosmids have been sequenced together with two PCR products and a subcloned DNA fragment derived from a cosmid identified as pXB564 in order to generate the plasmid's complete nucleotide sequence.

After its assembly, the analysis of the entire nucleotide sequence of pNGR234a, especially the determination of putative coding regions and the prediction of their expressible proteins and putative functions, was performed. Initially, analysis of the region covered by cosmid pXB296 was extended to cosmids pXB368 and pXB110. Thus, in approximately 100 kb of the plasmid (position 417,796 - 517,279) most ORFs and their deduced proteins with different putative functions were predicted. Subsequently, the rest of pNGR234a was analyzed.

SUMMARY OF THE INVENTION

The present invention provides the complete nucleotide sequence of symbiotic plasmid pNGR234a or degenerate variants thereof of *Rhizobium* sp. NGR234.

The present invention also contemplates sequence variants of the plasmid pNGR234a altered by mutation, deletion or insertion.

Also encompassed by the present invention are each of the ORFs derivable from the nucleotide sequence of pNGR234a or variants thereof.

In a preferred embodiment, the ORFs derived from the nucleotide sequence of pNGR234a encode the functions of nitrogen fixation, nodulation, transportation, permeation, synthesis and modification of surface poly- or oligosaccharides, lipo-oligosaccharides or secreted oligosaccharide derivatives, secretion (of proteins or other biomolecules), transcriptional regulation or DNA-binding, peptidolysis or proteolysis, transposition or integration, plasmid stability, plasmid replication or conjugal plasmid transfer, stress response (such as heat shock, cold shock or osmotic shock), chemotaxis, electron transfer, synthesis of isoprenoid compounds, synthesis of cell wall components, rhizopine metabolism, synthesis and utilization of amino acids, rhizopines, amino acid derivatives or other biomolecules, degradation of xenobiotic compounds, or encode proteins exhibiting similarities to proteins of amino acid metabolism or related ORFs, or enzymes (such as oxidoreductase, transferase, hydrolase, lyase, isomerase or ligase).

In another preferred embodiment, the ORFs are under the control of their natural regulatory elements or under the control of analogues to such natural regulatory elements.

The present invention also provides the sequences of the intergenic regions of pNGR234a which, in a preferred embodiment, are regulatory DNA sequences or repeated elements. In a further preferred embodiment, the intergenic sequences are ORF-fragments.

Also provided by the present invention are mobile elements (insertion elements or mosaic elements) derivable from the nucleotide sequences of the present invention.

The present invention also contemplates the use of the disclosed nucleotide sequences or ORFs in the analysis of genome structure, organisation or dynamics.

Also provided by the present invention is the use of the nucleotide sequences or ORFs in the subcloning of new nucleotide sequences. In a preferred embodiment, the new nucleotide sequences are coding sequences or non-coding sequences.

In yet a further preferred embodiment, the nucleotide sequences or ORFs are used in genome analysis and subcloning methods as oligonucleotide primers or hybridization probes.

The present invention further provides proteins expressible from the disclosed nucleotide sequences or ORFs.

Also contemplated by the present invention is the use of the disclosed nucleotide sequences, individual ORFs or groups of ORFs or the proteins expressible therefrom in the identification and classification of organisms and their genetic information, the identification and characterisation of nucleotide sequences, the identification and characterisation of amino acid sequences or proteins, the transportation of compounds to and from an organism which is host to said nucleotide sequences, ORFs or proteins, the degradation and/or metabolism of organic, inorganic, natural

09939964-082704
T02280-4965E650

5 The present invention also provides plasmid pNGR234a of *Rhizobium* sp. NGR234 comprising the disclosed nucleotide sequence or any degenerate variant thereof.

The plasmids of the invention may be produced recombinantly and/or by mutation, deletion, insertion or
15 inactivation of an ORF, ORFs or groups of ORFs.

30 The nucleotide sequences of the present invention were advantageously obtained using known cycle sequencing methods. The preferred dye terminator/thermostable sequenase shotgun sequencing method used to generate the nucleotide sequences of the present invention, when applied
35 to cosmid and when compared to other sequencing methods, was shown to yield sequence reads of the highest fidelity. Consequently, the speed of assembly of particular cosmid was increased, and the resultant high-quality sequences

required little editing or proofreading. Thus, the preferred sequencing method described herein was successfully used to generate the complete nucleotide sequence of all the overlapping cosmids of plasmid pNGR234a, thereby resulting in the assembly of the complete sequence of the plasmid.

The complete sequence of pNGR234a is disclosed for the first time in this application, as are the majority of the ORFs predicted within the sequence. Putative functions have been ascribed to the novel and inventive ORFs disclosed herein and the proteins for which they code.

15 BRIEF DESCRIPTION OF DRAWINGS

The present invention is described below and illustrated thereafter in the appended examples, with reference to the following figures:

20

Figure 1 A comparative graph showing the comparison of sequences from pXB296 created by different cycle sequencing methods.

25 **Figure 2** A schematic diagram showing the organization of the predicted ORFs in pXB296 from *Rhizobium* sp. NGR234.

30 **Figure 3** The complete nucleotide sequence of plasmid pNGR234a (with the pages labelled sequentially from 19961 to 1996142).

35 **Figure 4** A schematic diagram showing the map of the 20 sequenced cosmids covering the 536 kb symbiotic plasmid pNGR234a of *Rhizobium* sp. NGR234.

Figure 5 A diagram indicating multiple alignments of the nucleotide sequence of the replication origins

0993964-082704
FOZ280-4956E560

of various plasmids.

Figure 6 A diagram indicating multiple DNA sequence alignments of the regions containing the origin of transfer of various plasmids.

Figure 7 A schematic diagram showing a circular representation of the symbiotic plasmid pNGR234a of NGR234.

DETAILED DESCRIPTION OF THE INVENTION AND BEST MODE

Comparison of Different Shotgun Sequencing Strategies

The following is a more detailed description of certain key aspects of the present invention.

GC-rich cosmids were examined to investigate whether they could be sequenced much more efficiently using dye terminators throughout the shotgun phase instead of dye primers. As a test case, cosmid pXB296 with a GC content of 58 mol% from pNGR234a, the symbiotic plasmid of *Rhizobium* sp. NGR234, was exclusively sequenced using dye terminators in combination with a thermostable sequenase [Thermo Sequenase (Amersham)]. Another rhizobial cosmid with identical GC content, pXB110, was sequenced using traditional dye primer chemistry and *Taq* DNA polymerase.

Using the dye terminator/thermostable sequenase shotgun strategy, it was shown that most, if not all, compressions could be resolved and reads were produced with the highest fidelity among all sequencing chemistries tested. As a result, a much faster assembly of cosmid pXB296 in comparison to pXB110 was obtained. The shotgun data could be assembled into a high-quality sequence without extensive editing and proofreading. By measuring the error rate in overlapping regions between individual cosmids from

pNGR234a, as well as the cosmid vector sequence itself (data not shown), it was estimated that the accuracy of the pXB296 sequence is higher than 99.98%. Using other thermostable sequenases such as AmpliTaq FS (Perkin-Elmer), similar results were expected because thermostable sequenases have similar properties.

Dye primer chemistry in combination with Thermo Sequenase was also examined. Although the peak uniformity of signals was much improved over dye primer/Tag DNA polymerase data, the number of compressions in GC-rich shotgun reads was not reduced significantly. Compressions in shotgun raw data enormously increase the overall effort of editing, proofreading, and finishing a cosmid as shown for pXB110 (Table 1).

Because of their longer reading potential, dye primer reads are helpful for gap closure. However, using ABI 373A sequencers (Applied Biosystems, Inc. (ABI), Perkin-Elmer, Foster City, CA, USA), dye primer reads are, on average, only ~50 bases longer than dye terminator reads.

Using the experimental conditions of the present invention, shotgun sequencing with dye terminators and a thermostable sequenase is superior because for GC-rich cosmid templates it removes most of the compressions and this leads to a several-fold improvement in assembling and finishing of cosmid-sized projects. Although dye terminators are slightly more expensive than dye primers, the overall saving in time for finishing projects has, in our experience, a much greater effect on general costs.

It has been shown that the strategy of the present invention is effective for high-throughput shotgun sequencing of GC-rich templates. This strategy was therefore used to sequence the remaining 19 overlapping cosmids of the symbiotic plasmid pNGR234a of *Rhizobium* sp. NGR234. In total, 20 cosmids, two PCR products (1.5 and

0093964-082701
10/23/90 14:56:50

Table 1. Comparison of the assembly of the sequence data from cosmids pXB296 (dye terminator shotgun reads) and pXB110 (dye primer shotgun reads)

Data assembly	pXB296	pXB110
Average length of the shotgun reads (bases)	332	378
No. of shotgun reads used for assembly	786	899
No. of shotgun reads assembled with 4% mismatch ^a	736	308
No. of shotgun reads assembled with 25% mismatch ^a	775	879
No. of contigs ^b longer than 1 kbp	3	25
No. of contigs left after editing ^c	2	4
No. of additional reads (gap closure and proofreading) ^d	32	191
Total length of cosmid insert (bp)	34,010	34,573
Sequencing redundancy (per-bp)	8.0	10.5

^aAssembling program: XGAP; principal autoassembling conditions: normal shotgun assembly, joins permitted, minimum initial match = 15, maximum no. of pads per reading during the alignment procedure = 8, maximum no. of pads per reading in contig to align any new reading = 8, alignment mismatches 4% and 25%, respectively.

^bContiguous parts of sequence created by overlapping reads.

^cLengths of contigs: 6–10 kbp (pXB296); 2–12 kbp (pXB110).

^dReads necessary for closing gaps and making single-stranded regions double-stranded by primer walking on selected templates and, in case of pXB110, for solving ambiguities (compressions) by the resequencing of clones with universal primer and dye terminators.

2.0 kb in length) and a 1.5 kb restriction fragment were sequenced in order to generate the complete pNGR234a sequence (Figure 4).

5 Genetic Organization of pXB296

All 28 predicted open reading frames (ORFs) in pXB296 (Figure 2) show significant homologies to database entries (Table 2). The first putative gene cluster (cluster I) containing ORF1 to ORF5 corresponds to various oligopeptide permease operons (Hiles et al., 1987; Perego et al., 1990). Only ORF5 shows homology to a gene from a different bacterium, *Bacillus anthracis* (Makino et al., 1989). Each homologue encodes membrane-bound or membrane-associated proteins suggesting that all five ORFs are involved in oligopeptide permeation.

Organization of the predicted gene cluster IV, including the *nifA* homologue ORF16 (*fixABCX*, *nifA*, *nifB*, *fdxN*, ORF, *fixU* homologues, position 16,746 - 24,731), the predicted locations of the σ^{54} -dependent promoters and the *nifA* upstream activator sequences (Figure 2), correspond to the organization found in *Rhizobium meliloti* and *Rhizobium leguminosarum* bv. *trifolii*. (Iismaa et al., 1989; Fischer, 1994). *NifA* is a positive transcriptional activator (Buikema et al., 1985), whereas *nif* and *fix* genes are essential for symbiotic nitrogen fixation. Identification of σ^{54} -dependent promoter sequences, together with the upstream activator motifs upstream of ORF21, ORF22, and ORF23, suggests that these ORFs may play an important, but still undefined, role in symbiosis.

Inevitably, large-scale sequencing uncovers differences with already published sequences. van Slooten et al. (1992) cloned a 5.8 kb *EcoRI* fragment from *Rhizobium* sp. NGR234 and sequenced 2067 bp by manual radioactive methods (EMBL accession no. S38912). This sequence exhibits 2.4% mismatches with the corresponding sequence in pXB296.

Table 2. Putative ORFs of pXB296 and homologues of the deduced amino acid sequences to known proteins

ORF ^a	st. ^b	position on cosmid (base no.) ^c	ribosomal binding site: SD-sequence - distance from start codon (bases)- start codon ^d	no. of deduced amino acids	homologous amino acids (position)	homologous protein	name	length (aa) ^e	function ^f	accession no.	identity (%) ^g	similarity (%) ^g
SD-sequence: 5'-TAAGGAGGTGA-3'												
ORF1 ^h	+	00001-00625		>207	1-207	OppB	OppB	306	oligopeptide	X05491	45	68
ORF2	+	00628-01503	GTATCCGGT-7-ATG	291	2-289	OppC	OppC	305	permease	X56347	37	63
ORF3	+	01503-02512	AGCGGAGG-7-ATG	335	8-327	OppD	OppD	336	proteins	X56347	49	69
ORF4	+	02509-03570	TGAAGTGGT-6-ATG	353	2-323	OppF	OppF	334		X05491	51	69
ORF5	+	03606-04991	CAAGGA-6-ATG	461	1-458	CapA	CapA	411	encapsulation protein	M24150	25	48
ORF6	+	05460-06863	CCGAGAGG-8-ATG	467	1-464	BioA	BioA	455	aminotransferase	M29292	29	55
ORF7	+	06888-08426	GCCTTCGG-5-GTG	512	97-509 34-510	ORF ⁱ GapD		417 482	unknown succinic semialdehyde dehydrogenase	D37877 M38417	36 33	58 57
ORF8	-	09781-10860	GAACGTGG-8-ATG	359	72-299	ORF ^j	ORF ^j	414	transposase homologue minicircle DNA	X15942	30	48
ORF9	+	11124-12455	7-7-ATG	443	2-443	GLUD1	GLUD1	558	glutamate dehydrogenase	M37154	41	60

Table 2. (Continued)

ORF10	-	13370-14116	AAAGGA-6-ATG	248	1-245	ORF1	231	transposase	X79443	45	64
ORF11	-	14128-15672	CATGGAG-7-TTG	514	1-513	ORF1	558	homologues, IS1162	X79443	41	62
ORF12	-	16712-16942	GAAGGA-8-ATG	76	1-70	FixU	70	unknown	P42710	63	80
ORF13	-	16939-17265	ACAAGAGG-7-ATG	109	1-79 15-107	ORF2 ¹ NiZ	>78 159	unknown involved in FeMo- cofactor synthesis	X07567 M20568	53 39	81 56
ORF14	-	17349-17543	CCAGGAG-9-ATG	64	1-64	FixN	64	ferredoxin-like	M21841	80	87
ORF15	-	17585-19066	AGTGGAG-7-ATG	493	1-493	NiIB	490	involved in FeMo- cofactor synthesis	M15544	73	84
ORF16	-	19292-20962	ATTGG-12-ATG	556	9-556	NiIA	541	transcriptional regulator	X02615	59	72
ORF17	-	21129-21422	AGGGGAG-7-ATG	97	1-97	FixX	98	required for	M15546	84	87
ORF18	-	21437-22744	AACTEAGGT-7-ATG	435	1-435	FixC	435	nitrogen	M15546	83	90
ORF19	-	22755-23864	ATAGGAG-6-ATG	369	18-369	FixB	353	fixation	M15546	79	89
ORF20	-	23874-24731	TAAAGAG-5-ATG	285	1-285	FixA	292		M15546	74	85
ORF21	-	25148-25468	CCAGGAG-10-ATG	106	1-106	ORF118 ¹	108	unknown	X13691	55	71
ORF22	-	26145-26711	GAAGGAG-9-ATG	188	9-199 1-173	-	241 166	hypothetical protein peroxisomal protein	U32739 U11244	47 32	64 57

Table 2. (Continued)

ORF23	+	27169-27861	GAGGGA-7-ATG	230	1-167	NifQ	167	probably involved in Mo-processing	X13303	37	57
ORF24	+	27920-29434	CTGGGAGG-18-ATG	504	1-454 8-454	DctA1 DctA2	456 449	C ₄ -dicarboxylate transporter	S38912 S38912	97	98
ORF25	+	29431-30675	TTCGGCGG-12-ATG	414	2-414	CamC	415	cyp450-like	M12546	34	53
ORF26	+	30676-31332	TTCGGG-5-TTG	218	30-190	LiaA	155	γ -hexachloro-cyclohexan-dechlorinase	D90355	27	51
ORF27	+	31329-33035	ACTGGAG-10-ATG	568	28-270	FabG	244	reductase	M84991	38	57
ORF28 ^k	+	33173-34010	CAAGGAG-5-ATG	>279	294-534 1-279	LuxA	355	luciferase α -subunit	M10961	23	49

^a(ORF) Open reading frame.

^b(st.) Plus or minus strand.

^cPosition on cosmid: from the first base of the start codon to the last base of the stop codon; alternative start points are 6912/6927/7017 (ORF7), 10665/10656 (ORF8), 11220 (ORF9), 15699/15651 (ORF11), 17322/17271 (ORF13), 20995/21076 (ORF16), 26744 (ORF22), 27229/27304 (ORF23), 27941 (ORF24), and 30751/30754 (ORF26).

^d(SD sequence) Shine-Dalgarno sequence (Shine and Dalgarno 1974). Bases underlined are identical with the Shine-Dalgarno sequence. The following possible start codons were considered: ATG, GTG, or TTG.

^e(aa) Amino acids.

^fOrganisms: *Salmonella typhimurium*, *Bacillus subtilis* (OppBCDF), *Bacillus anthracis* (CapA), *Bacillus sphaericus* (BioA), *Streptomyces hygroscopicus* (ORF7 homolog), *Escherichia coli* (GapD), *Streptomyces coelicolor* (ORF8 homolog), *Homo sapiens* (GLUD1), *Pseudomonas fluorescens* (ORF10, ORF11 homologs), *Rhizobium leguminosarum* (FixU), *Rhodobacter capsulatus* (ORF13 homolog), *Azotobacter vinelandii* (NifZ), *Rhizobium meliloti* (FdxN, NifBA, FixXCB), *Bradyrhizobium japonicum* (ORF118), *Haemophilus influenzae* (hypothetical protein), *Lipomyces kononenkoae* (peroxisomal protein), *Klebsiella pneumoniae* (NifQ), *Rhizobium* sp. NGR234 (DctA), *Pseudomonas putida* (CamC), *Pseudomonas paucimobilis* (LiaA), *Escherichia coli* (FabG), *Vibrio harveyi* (LuxA).

^gIdentity and similarity were calculated using the program BESTFIT (Smith and Waterman 1981).

^h(ORF1) 3' end.

ⁱTranslated ORF.

^k(ORF28) 5' end.

It contains the gene *dctA* (encoding a C₄-dicarboxylate permease), which is 144 bases shorter than in pXB296. In this respect, a single nucleotide deletion in position 29,248 of the cosmid sequence close to the 3' end of the gene causes a frameshift leading to a DctA product extended by 48 residues. van Slooten et al. (1992) also failed to identify the *nifQ* homologue, ORF23 (position 27,169 - 27,861), presumably because they overlooked a small *Xho*I fragment located between positions 27,349 and 27,536 on pXB296. Expression studies allowed these investigators to define a putative σ^{54} -dependent promoter in a 1.7 kb *Sma*I fragment (position 27,094 - 28,818 in pXB296). This fragment stretches from the upstream region of ORF23 to the 5' part of *dctA*. The 58 bp intergenic region between ORF23 and *dctA* contains a stem-loop structure but no obvious promoter sequence. Possibly the promoter that controls *dctA* is located upstream of ORF23 (e.g. the minimal consensus sequence included in GGGGGCACAATTGC at position 27,098 - 27,111). Although clones containing *dctA* complemented mutants of *R. meliloti* and *R. leguminosarum* for growth on dicarboxylates, the growth of the NGR234 *dctA* deletion mutant was not affected (van Slooten et al., 1992). Nevertheless, this mutant was unable to fix nitrogen in nodules. Because *dctA* is now possibly part of a larger transcription unit, the symbiotic phenotype may also result from the inactivation of downstream genes.

Interestingly, the GC content of the predicted pXB296 ORFs ranges from 53.3 mol% to 64.6 mol%, with an overall cosmid GC content of 58.5 mol%. Genomes of *Azorhizobium*, *Bradyrhizobium*, and *Rhizobium* species have GC contents of 59 mol% to 65 mol% (Padmanabhan et al., 1990), with 62 mol% reported for *Rhizobium* sp. NGR234 (Broughton et al., 1972). Although pXB296 covers <7% of the complete symbiotic plasmid sequence, its lower overall GC value suggests that symbiotic genes might have evolved by lateral transfer from other organisms. In this case, methods of the type applied in the present invention will become even more relevant in



5

Extending the analysis of pXB296 to a 100 kb region stretching from position 417,796 to 517,279 on the symbiotic plasmid pNGR234a led initially to the assignation of only 76 ORFs listed within Table 3 (excluding the first incomplete ORF noted in the analysis of pXB296 ("ORF1" of Table 2)). The ORFs y4tQ to y4vJ (excluding ORFs y4uD and y4uG and excluding ORF-fragments fu1, fu2, fu3, fu4 and fv1; see Table 3) are identical to the ORFs 2 to 28 of the analysis of pXB296 in Table 2 apart from minor revisions (N.B. the analysis recited in Table 3 should be taken as the definitive analysis - Table 2 merely represents preliminary findings). The cosmid pXB110, which was sequenced with the dye primer shotgun sequencing strategy in order to compare it with the dye terminator shotgun sequencing strategy used to sequence cosmid pXB296, in combination with pXB296 and pXB368 cover nearly this entire region. A PCR product and a restriction fragment of cosmid pXB564 also had to be sequenced in order to fill in the gap from position 480,607 to 483,991 between cosmids pXB368 and pXB110 (Figure 4). Among the 76 predicted ORFs, 7 ORFs and their deduced proteins show no homologies to database entries. The other predicted ORFs and their deduced proteins do exhibit such homologies and therefore play putative roles in nitrogen fixation (ORFs y4uJ to y4vB, y4vE, y4vN to y4vR, y4wK and y4wL), nodulation (ORFs y4yC and y4yH), transportation (ORFs y4tQ to y4uA, y4vF and y4wM), secretion of proteins or other biomolecules (ORFs y4yI and y4yO), transcriptional regulation/DNA binding (ORFs y4wC and y4xI), in amino acid metabolism or metabolism of amino acid derivatives (ORFs y4uB, y4uC, y4uF, y4wD, y4wE and y4xN to y4yA), degradation of xenobiotic compounds (ORFs y4vG to y4vI), in peptidolysis/proteolysis (ORFs y4wA and y4wB) or transposition (ORFs y4uE, y4uH and y4uI) (see Table 3). The

Table 3 : List of the predicted functional ORFs and of fragments representing putative remnants of functional ORFs

ORF ^a	func-tional name	st. ^b	position in plasmid (base no.) ^c	no. of deduced amino acids	hom. amino acids (position)	hom. protein			I/ % ^f	S/ % ^f	note ^g
						name	length (aa) ^d	accession no. ^e			
y4aA		-2/3	534696 - 000474	647	16-646	Shc	658	X86552	78	88	prob. squalene-hopene-cyclase; put. operon y4aABCD; inv. in synthesis of an isoprenoid compound
y4aB		-3	000523 - 001776	417	6-415	ORF1	414	X80766	43	63	put. flavoprotein oxidoreductase
y4aC		-2	001776 - 002615	279	3-247	Psy1	419	X68017	34	50	put. phytoene synthase
y4aD		-1	002612 - 003490	292	10-195	Ctrl	342	L37405	33	51	hyp. protein hom. to squalene and phytoene synthetases
fa1		-3	003487 - 004011								fragmentous character
y4aF	noK	-3	005173 - 006117	314	9-310	ORF14.8	321	U46859	51	70	put. NAD-dep. nucleotide sugar epimerase/dehydrogenase; NoeJL/NoeZ/NoK inv. in biosynthesis of fucose moiety of Nod factors
y4aG	noeH	-2	006126 - 007181	351	4-339	RfbD	348	U24571	65	80	put. GDP-D-mannose dehydratase
y4aH	nodZ	-1	007426 - 008394	322	3-254	NodZ	324	L22756	69	83	put. fucosyltransferase
y4aI	noeK	-3	008623 - 010047	474	5-471	ORF5	483	U47057	42	59	put. phosphomannomutase
y4aJ	noeJ	+3	010110 - 011648	512	33-498	XanB	466	M83231	50	65	put. mannose-1-phosphate guanylyltransferase
y4aK		+2	012125 - 012277	50							hyp. 5.5 kd protein
y4aL	nodDI	+2	012380 - 013348	322	1-322 1-310	NodDI NodDI2	322 312	Y00059 this work	98 68	99 84	transcriptional regulator (LysR family); high similarity to Y4xH(NodD2)
y4aM		+3	013911 - 014342	143	7-132 1-143	ORF3 Y4wC	127 143	L13845 this work	50 69	66 77	put. DNA-binding protein; high similarity to Y4wC
y4aN		+1	014488 - 014934	148	1-129	ORF3	128	X04833	41	56	homologue located nearby the replicator region of pRI44b
y4aO		+3	015065 - 015643	192							hyp. 21.8 kd protein; low similarity to Y4nF(<30% id.)
y4aP	mucR	+3	016161 - 016592	143	1-143	MucR	143	L37353	89	95	put. transcriptional regulator (Ros/MucR family); exopolysaccharide synthesis
y4aQ		-2	017016 - 017582	188	15-167	NoI265	266	X74068	33	50	hyp. 20.4 kd protein; similar to Y4hP, Y4jD, Y4qI
y4aR		+2	017798 - 018121	107							hyp. 12.1 kd protein
y4aS		+1	018121 - 018666	181							hyp. 20 kd protein
fa2		+3	018912 - 019664	250	126-250	Tap	465	U04047	38	51	hyp. protein fragment
					78-150 3-266	Y4iG Y4bF	90 457	this work this work	93 53	97 73	

y4bA		-2	019674 - 021758	694		1-393 406-532 532-694	fo6 fo5 fo4	430 136 143	this work this work this work	89 83 77	95 94 83	hyp. 78.7 kd protein; identical to Y4pH
y4bB		-3	021748 - 022014	88		2-88	Y4oL	88	this work	63	69	hyp. 9.7 kd protein precursor; identical to Y4pl
y4bC		-1	022034 - 022483	149		1-149	Y4oM	149	this work	79	88	hyp. 16.8 kd protein; identical to Y4pl
y4bD		-2	022674 - 022943	89		20-89	Y4oN	70	this work	73	84	hyp. 10.2 kd protein; identical to Y4pK
fbI		+2	022985 - 023659	224		36-224	Y4bF	457	this work	42	63	hyp. protein fragment
y4bF		+1	023953 - 025326	457		130-436	Tnp	465	U04047	31	46	put. transposase; upstream of this ORF (23875-23987) 89% nt-id. to part of origin of replication-region (<i>R. meliloti</i> , S66221)
y4bG		+1	025870 - 026685	271								hyp. 30 kd protein precursor
y4bH		+1	028513 - 028788	91								hyp. 9.6 kd integral membrane protein
y4bI		+3	028860 - 029276	138		3-108	H11631	190	U00085	41	61	hyp. 15.3 kd protein precursor
y4bJ		+1	029392 - 031284	630		429-564	HtrA	503	L20127	40	53	hyp. 67.9 kd integral membrane protein, distantly related to peptidase family S2C
y4bK		+2	031625 - 032293	222		83-212	ORF1	215	D84146	25	45	hyp. 24.3 kd protein
y4bL		+1	032641 - 034191	516		7-515	ORF1	558	X79443	44	63	identical to Y4kJ and Y4tB; similar to Fo3 and Fo7; put. transposase
y4bM		+3	034188 - 034979	263		6-516	Y4uI	515	this work	48	66	identical to Y4kJ and Y4tA; put. insertion sequence ATP- binding protein; similarity to Y4pL, Y4uH, also to Y4sD/Y4nD/Y4tQ
y4bN		+1	035278 - 036573	431		1-203	ORF2	231	X79443	45	62	
y4bO		+1	036646 - 038466	606		6-248	Y4pL	245	this work	55	73	
y4cA		-1	038576 - 042169	1197		6-254	Y4uH	248	this work	48	68	
y4cB		-3	042226 - 042522	98		1-263	Y4tQ	298	this work	31	56	
y4cC		-3	042556 - 044109	517								hyp. 137.7 kd protein; largest protein in pNGR234a
y4cD		-2	044106 - 046028	640								hyp. 10.2 kd integral membrane protein
y4cE		-3	046486 - 047661	391								hyp. 57.8 kd protein
y4cF		-1	047687 - 048829	380								hyp. 71.6 kd protein
y4cG		+2	049361 - 050278	305		16-173 17-222	Pin Y4tS	184 183	K00676 this work	50 40	68 60	hyp. 43.4 kd protein hyp. 41.8 kd protein prob. DNA invertase "resolvase-type"
y4cH		-2	050427 - 050636	69		4-65	CspS	70	L23115	56	70	prob. cold shock regulator

y4cI		-2	053202 - 054416	404	1-397	RepC	405	X04833	60	73	put. replication protein C
y4cJ		-3	054571 - 055551	326	1-317	RepB	319	X89447	39	55	put. replication protein B
y4cK		-2	055608 - 056831	407	10-404	RepA	398	X89447	58	73	put. replication protein A
y4cL	<i>tral</i>	+2	057635 - 058261	208	1-206	Tral	212	U43675	55	66	prob. autoinducer synthetase (inv. in control of conjugal transfer)
y4cM	<i>trbB</i>	+3	058272 - 059249	325	3-325	TrbB	323	U43675	80	88	prob. conjugal transfer protein (PuIE family)
y4cN	<i>trbC</i>	+1	059239 - 059622	127	1-115	Y4oG	125	this work	25	51	
y4cO	<i>trbD</i>	+2	059615 - 059914	99	7-127	TrbC	134	U43675	69	78	prob. conjugal transfer protein (integral membrane prot.)
y4cP	<i>trbEa</i>	+3	059925 - 060374	149	1-99	TrbD	99	U43675	70	89	prob. conjugal transfer protein (integral membrane prot.)
y4cQ	<i>trbEb</i>	+1	060394 - 062382	662	1-136	TrbE	820	U43675	80	91	prob. conjugal transfer protein (hom. to 5' part of <i>trbE</i>)
y4dA	<i>trbJ</i>	+2	062354 - 063157	267	5-659	TrbE	820	U43675	83	90	prob. conjugal transfer protein (hom. to 3' part of <i>trbE</i>)
y4dB	<i>trbK</i>	+1	063154 - 063351	65	1-107	TrbJ	175	U43675	60	69	prob. conjugal transfer protein
y4dC	<i>trbL</i>	+3	063345 - 064520	391	194-267				71	79	
y4dD	<i>trbF</i>	+2	064544 - 065206	220	5-65	TrbK	75	U43675	40	56	prob. conjugal transfer protein precursor
y4dE	<i>trbG</i>	+1	065224 - 066036	270	3-387	TrbL	395	U43675	74	85	prob. conjugal transfer protein (integral membrane prot.)
y4dF	<i>trbH</i>	+1	066040 - 066486	148	1-220	TrbF	220	U43675	80	90	prob. conjugal transfer protein
y4dG	<i>trbI</i>	+3	066498 - 067793	431	6-270	TrbG	284	U43675	74	84	prob. conjugal transfer protein precursor
y4dH	<i>traR</i>	+2	068096 - 68806	236	1-147	TrbH	159	U43675	55	68	prob. conjugal transfer protein precursor (with lipid anchor)
y4dI	<i>traM</i>	-1	068810 - 069133	107	1-430	TrbI	433	U43675	66	79	prob. conjugal transfer protein (integral membrane prot.)
y4dJ		+3	069351 - 069584	77	7-236	TraR	234	Z15003	28	45	prob. transcriptional activator of conjugal transfer genes (LuxR family)
y4dK	<i>fdI</i>	-1	069629 - 069949	106	8-101	TraM	102	U43674	30	51	prob. modulator of TraR/autoinducer-mediated activation of <i>tra</i> genes
y4dL		-2	069936 - 070250	196	1-67	ORF	84	X16458	37	59	hyp. transcriptional regulator (PbsX family); low similarity to N-terminus of Y4dL
y4dM		+1	070603 - 071193	409	(2-85)	ORFA	400	X67861	39	58	hyp. 11.8 kd protein
y4dN		+2	071186 - 072415	409	-						put. transposase fragment
y4dO		+1	072787 - 072975	62	1-357	HipA	440	M61242	31	46	hyp. 21.8 kd protein; low similarity to Y4dJ
y4dP		-1	073550 - 073951	133	3-405	Y4mE	420	this work	34	56	hyp. 45.3 kd protein; homolog affects frequency of persistence after inhibition of cell wall or DNA synthesis
y4dQ		-1	074423 - 075025	200	-						hyp. 7 kd protein
y4dR	<i>traB</i>	-2	075042 - 076205	387	12-121	ORF	381	D83536	43	57	hyp. 14.9 kd (fragmentous?) protein; homology to intron protein of <i>P. anserina</i> continues in fr.-2 (73541-73467)
y4dS	<i>traF</i>	-3	076195 - 076761	188	1-48	ORFR2	57	U43674	72	89	hyp. 21 kd protein; hom. to conjugal transfer region 1
					56-198	ORFR3	154		47	71	
					1-387	TraB	421	U40389	61	72	prob. conjugal transfer protein
					20-188	TraF	176	U40389	55	73	prob. conjugal transfer protein

y4dS		traA	-2	076758 - 080066	1102	1-1102	TraA	1100	U43674	67	79	prob. conjugal transfer protein (relaxase)
y4dT		traC	+3	080319 - 080627	102	1-102	TraC	98	U40389	64	80	prob. conjugal transfer protein
y4dU		traD	+1	080632 - 080847	71	1-71	TraD	71	U43674	77	84	prob. conjugal transfer protein
y4dV		traG	+2	080834 - 082756	640	1-631	TraG	658	U40389	71	83	prob. conjugal transfer protein
fd2			+	083002 - 083293			ORFL1	152	U43674			fragments hom. to ORFL1 (conjugal transfer region1); frameshifts: 83072 (1>3), 83161 (3>2)
y4dW			+1	083305 - 083919	204							hypothetical 22.9 kd protein
y4dX			+1	083944 - 84522	192							hypothetical 20.6 kd protein
y4eA			-2	084570 - 084836	88							hypothetical 9.9 kd protein
y4eB			-3	084976 - 085290	104							hypothetical 11.6 kd protein
fe1			-	085829 - 088007			MerA	474	X65467			put. fragments; homology to mercuric reductase, put. frameshifts: 86592 (-1<3), 87288 (-3<2)
y4eC			-2	088305 - 089228	307	14-306	TraC-1	1061	X59793	38	55	hyp. 34.2 kd protein; hom. to 5'end. of traC-1 from plasmid RP4
y4eD			+1	091051 - 092178	375	51-136	ORF145	145	X52594	29	55	put. phosphodiesterase; low homology to glycerophosphoryl-diester-phosphodiesterase
y4eE			+1	092212 - 093288	358							hyp. 38.5 kd protein
fe2			-	093572 - 093969			TnpA		U14952			fragments of put. transposase; put. frameshift: 93798 (2<3)
y4eF			-1	093980 - 094735	251	2-236 1-251	Int Y4qK	259 308	U14952 this work	37 92	53 94	put. integrase/recombinase ("phage-type"); similar to Y4rF (35% aa-id.); low similarity to Y4rABCDE
fe5			-1	094988 - 095188	66	1-66 1-66	Fq6 Y4rC	66 332	this work this work	79 41	94 55	put. defective integrase/recombinase
fe3			-	095343 - 096025			Int	259	U14952			fragments hom. to integrase; put. frameshift: 95559-95671 (-2<-1)
y4eH		noIL	-2	096093 - 097193	366	11-359	NoIL	373	U22899	63	77	modulation protein; hyp. acetyl transferase
y4eI			-2	097914 - 098225	103							hyp. 11.1 kd protein with transmembrane domain
fe6			+3	098358 - 098657	99	3-98	AatB	410	L12149	40	55	hyp. 10.3 kd protein fragment, hom. to C-terminal part of bacterial aminotransferases
y4eK			+2	098675 - 099421	248	10-245	Adh	252	U00084	37	53	hyp. short chain type dehydrogenase/reductase
y4eL			+3	099447 - 100193	248	1-244	Gno	256	X80019	31	47	hyp. short chain type dehydrogenase/reductase
fe4			+	100270 - 101901			IivG		M37337			put. fragment; put. frameshifts: 100721 (1>2), 101728 (2>1)
fe7			-1	101585 - 102298	237	1-103	Tnp	398	U08627	91	95	put. truncated transposase-like protein; similar to Y4pO
y4eN			-3	102625 - 102936	103							hyp. 11.5 kd protein
y4eO			-2	102933 - 103598	221							hyp. 24.5 kd protein
y4fA			-1	103805 - 106342	845	327-837 7-845	McpA Y4sI	657 756	X66502 this work	41 29	59 49	prob. methyl-accepting chemotaxis protein

y4IB		+3	106620 - 108614	664														hyp. 73.7 kd protein
y4IC		+3	109884 - 110618	244	10-163	DszA	453											hyp. (fragmentous?) monooxygenase; extended homology to DszA in fr.2: 110372 to 110506.
y4ID		-1	110516 - 111178	220														hyp. 24.6 kd integral membrane protein
y4IE		-2	111195 - 111677	160														hyp. 17.2 kd protein precursor
y4IF		-1	111803 - 112348	181														hyp. 19.5 kd protein
y4IG		-2	112338 - 112727	129														hyp. 14.5 kd protein
y4IH		-1	113474 - 113782	102														hyp. 11.6 kd protein
flI		-3	113779 - 114114	111	61-97	DppF	330											hyp. protein fragment, similar to central region of oligo/di-peptide ABC transporter ATP-binding proteins
y4II		-2	114348 - 115379	343	3-210	RopA	318											put. outer membrane protein (porin) precursor
y4IK		-2	116112 - 117395	427	275-421	XylS2	157											put. transcriptional regulator (AraC family)
y4IL		-3	117385 - 118212	275	9-243	ORF	268											hyp. 29.1 kd integral membrane protein, belongs to the inositol monophosphatase family
y4IM		-2	118209 - 119144	311														hyp. 35.5 kd protein
y4IN		-2	119145 - 120854	569	11-513	CysU	550											prob. ABC transporter permease protein; put. part of binding-protein-dependent transport system Y4INOP
y4IO		-1	120851 - 121870	339	12-247	PotA	381											prob. ABC transporter ATP-binding protein
y4IP		-1	121883 - 122959	358	32-293	SufA	338											prob. ABC transporter periplasmic binding protein precursor
y4IQ		+1	123016 - 124194	392	9-234	NagC	406											hyp. 41.6 kd protein; belongs to "ROK" family (transcriptional regulator or transferase)
y4IR		+1	124813 - 126453	546	88-539	IpaH	532											hyp. 60.5 kd protein, hom. to invasion plasmid antigen H
y4IA		-1	126806 - 127369	187														hyp. 20.9 kd protein; low similarity to Y4rE
y4IB		-2	127485 - 127904	139														hyp. 16.1 kd protein
y4IC		-1	127901 - 128479	192	1-178	ORF2	415											put. integrase/recombinase ("phage-type")
y4ID		-1	128579 - 128857	92														hyp. 10.5 kd protein
y4IE		+2	131021 - 131767	248														hyp. (fragmentous?) 27.7 kd protein; put. frameshifts: 131532 (2>1), 131892 (1>2)
y4IF		+2	132734 - 133786	350	4-345	RhsB	333											prob. dTDP-D-glucose-4,6-dehydratase (Y4FGH inv. in dTDP-L-rhamnose biosynthesis)
y4IG		+2	133790 - 134680	296	1-290	RhsD	288											prob. dTDP-4-dehydrothamnose reductase
y4IH		+1	134677 - 135537	286	2-285	RfbA	293											prob. glucose-1-phosphate thymidyltransferase
y4II		+3	135534 - 138263	909	276-894	RfbC	1275											hyp. 102.8 kd protein (homolog is involved in O-antigen biosynthesis)
y4IJ		-1	138737 - 139315	192														hyp. 21.1 kd protein
y4IK	fixF	+3	142026 - 143234	402	114-184 203-362	KpsS	389											necessary for functional nitrogen fixation, hom. to capsule polysaccharide export protein

y4gL		-3	143473- 144060	195	24-192	RhsC	188	U51197	53	65	prob. dTDP-4-dehydrohamnose-3,5-epimerase (inv. in dTDP-L-rhamnose biosynthesis)
y4gM		-2	144147- 145907	586	26-581	MsbA	582	Z11796	32	56	prob. ABC transporter ATP-binding protein
y4gN		+2	146075- 147226	383	52-297	VirA	304	L08012	29	46	hyp. 45 kd protein
y4hA		-1	147455- 148558	367	7-362	ChaA	366	L28709	34	58	put. ionic transporter
y4hB	<i>noeE</i>	-3	148819- 150078	419	3-138 197-289	F42G9.8	359	U00051	32	49	nodulation protein (put. sulfate transferase)
y4hC	<i>noeG</i>	-3	151051- 151782	243	18-229	u0002kb	243	U00024	27	42	nodulation protein (unknown function)
y4hD	<i>noIO</i>	-1	151979- 154021	680	1-126 140-496	NolN NolO	127 358	L22756	70	83	inv. in O-carbamoylation of Nod factors (sim. to NodU)
y4hE	<i>nodJ</i>	-3	154120- 154908	262	5-261	NodJ	262	J03685	69	84	prob. ABC transporter permease (see <i>nodI</i>)
y4hF	<i>nodI</i>	-3	154912- 155943	343	15-343	NodI	339	X55795	69	85	prob. ABC transporter ATP-binding transport protein; put. role: together with NodJ export of modified beta-1,4-N-glucosamine oligosaccharides
y4hG	<i>nodC</i>	-1	156095 - 157336	413	1-413	NodC	413	X73362	99	100	N-acetylglucosaminyltransferase
y4hH	<i>nodB</i>	-3	157351 - 157998	215	1-215	NodB	214	X73362	99	99	chitoooligosaccharide deacetylase
y4hI	<i>nodA</i>	-2	157995 - 158585	196	1-196	NodA	196	X73362	100	100	N-acyltransferase; <i>nodABC</i> involved in synthesis of backbone of modified N-acylated glucosamine oligosaccharides
y4hJ		-1	158993 - 159775	260	59-240	ORF2	251	L133618	68	81	hom. to part of coproporphyrinogen III oxidase (lacks C-terminus and conserved N-term. domain)
y4hK		+3	160722 - 161465	247							hyp. 25.4 kd integral membrane protein
y4hL		+1	161569 - 161826	85							hyp. 9.6 kd protein
y4hM		+1	163042 - 164253	403	53-169	Gfor	439	M97379	31	54	hyp. 43.9 kd protein (partially hom. to glucose-fructose oxidoreductase)
y4hN		+2	164600 - 165034	144	10-144	ORFA	135	X84099	38	53	hyp. 16 kd protein; partially hom. to Y4jB and Y4rG
y4hO		+1	165037 - 165384	115	1-115 1-115 1-115	ORF140 ORFC Y4jC	140 144 117	X74068 X84099 this work	100 54 36	100 69 62	hyp. 12.8 kd protein
y4hP		+1	165430 - 167088	552	1-215 80-328 362-492	no1265 ORF2 ORF3	266 258 163	X74068 M10204 M10204	97 67 47	97 79 61	hyp. 61.7 kd protein; similar to Y4aQ, Y4jD and Y4ql
y4hQ		+3	167091 - 167675	194	5-185 1-52	ORF3 ORF91	237 >91	X51418 X74068	35 96	53 98	hyp. 21.7 kd protein
y4hR		-3	167710 - 167934	74							hyp. 8.8 kd protein
f1		-	168208 - 168300								hyp. transposase fragment similar to <i>R. meliloti</i> ISRm2011-2
f2		+1	168430 - 168792	120	1-130 1-108	Y4iO Y4rJ	252 396	this work this work	78 74	87 87	put. defective transposase (homologous to N-terminal parts of Y4iO and Y4rJ)

f3		+2	168798 - 169190	130	1-109	ORF1A	317	M33159	37	55	put. defective transposase (hom. to C-terminal parts of Y4iO and Y4rJ); additionally weak homology to Y4pF/Y4sB and Y4qE (<30% identity)
y4iR		-3	169231 - 169716	161	15-145	PsiB	134	L26581	55	74	hyp. protein (homolog located in a polysaccharide biosynthesis inhibition operon)
y4iC		-2	169929 - 170621	230	58-123	ORF	161	Z73419	41	54	hyp. 25.8 kd protein (ORF=MTCY373.06)
y4iD		-3	170563 - 172551	662	137-342 418-605	ORF	495	Z73101	40	59	prob. monooxygenase (ORF=MTCY31.20)
y4iE		+3	173295 - 173702	135	1-135	Y4iL	155	this work	33	52	hyp. 15.4 kd (fragmentous?) protein; similar to Y4zA
y4iF		-3	174211 - 175128	305							hyp. 34.1 kd protein
y4iG		-2	175590 - 175862	90	1-73 1-73	Y4aT Y4bF	266 457	this work	93	97	hyp. 10.5 kd (fragmentous?) protein
y4iH		+2	176045 - 176764	239	1-236	Y4iT	336	this work	32	53	hyp. 26 kd protein precursor
y4iI		-2	176937 - 179048	703							hyp. 76.2 kd integral membrane protein
y4iJ		-2	179097 - 180887	596							hyp. 65.5 kd protein; low similarity to Y4iM
y4iK		-3	180940 - 181638	232							hyp. 26.8 kd protein; y4iKL: two fragments of one gene?; put. frameshift: 181884 (-3<-2)
y4iL		-2	181692 - 182990	432							hyp. 47.8 kd protein; y4iKL two fragments of one gene?; put. frameshift: 181884 (-3<-2)
y4iM		-2	183036 - 184334	432							hyp. 47.1 kd protein; low similarity to Y4iJ; y4iMN two fragments of one gene?; put. frameshift: 184440 (-2<-3)
y4iN		-3	184309 - 184935	208							hyp. 22.1 kd protein precursor; y4iMN two fragments of one gene?; put. frameshift: 184440 (-2<-3)
y4iO		-2	185679 - 186437	252	17-243	Thp	334	Z48244	29	46	put. transposase or transposase-fragment; additionally weak homology to Y4pF/Y4sB and Y4qE (<30% identity)
y4iP		-1	186437 - 186832	131	1-121	Fi2	120	this work	67	79	
y4iQ		-3	187162 - 188058	298	123-252 1-252	Fi3 Y4iJ	130 396	this work	78	87	hyp. 14.4 kd protein or fragment hom. to N-term. of Y4rJ identical to Y4nD/Y4sD; put. insertion sequence ATP-binding protein; similarity to Y4bM/Y4kI/Y4tA, Y4uH and weakly to Y4pL
y4jA		-2	188055 - 189569	504	4-163	Y4iJ	396	this work	58	80	identical to y4nE/y4sE; hyp. 57.2 kd protein with low similarity to IS21/IS408/IS1162 transposases
y4jB		+3	190248 - 190706	152	13-253 8-283 5-265 147-494 395-504 24-79	IsiA Fz4 ORF1	265 263 248 507 110 130	U38187 this work U38187 this work U19148	34 31 31 25 72 46	56 56 52 42 85 69	hyp. 16.7 kd protein; partially similarity to Y4hN; low similarity to Y4rG

y4JC		+2	190703 - 191056	117	1-115 1-117	ORFC Y4hO	144 115	X84099 this work	39 36	58 62	hyp. 13.1 kd protein; see y4hO
y4JD		+2	191105 - 192640	511	89-298 340-453 18-183	ORF2 ORF3 no1265	258 163 266	M10204 M10204 X74068	36 28 32	53 49 48	hyp. 56.7 kd protein; see y4hP
y4JE		+1	192637 - 193458	273							hypothetical (fragmentous?) 29.4 kd integral membrane protein; put. frameshift: 192996 (1>2; end of shifted ORF at 193183)
y4JF		-1	194771 - 196330	519							hyp. 55.4 kd integral membrane protein
y4JG		-3	196333 - 196821	162							hyp. 17.9 kd transmembrane protein
y4JH		-2	196818 - 197435	205							hyp. 23 kd protein
y4JI		-3	197428 - 197820	130							hyp. 13.6 kd protein
y4JJ		+1	198043 - 198300	85	1-85	StbC	103	L48985	67	76	put. plasmid stability protein
y4JK		+3	198297 - 198719	140	1-138	StbB	139	L48985	57	76	put. plasmid stability protein
y4JL		+3	199002 - 199664	220							hyp. 25.1 kd protein
y4JM		-2	199746 - 199958	70	11-58 15-58	Y4bF fb1	457 188	this work this work	75 50	79 64	hyp. 8 kd protein or protein fragment
y4JN		-3	199975 - 200415	146							hyp. 16.3 kd protein
y4JO		-3	201514 - 202479	321							hyp. 36.1 kd protein; y4JOP: two fragments of one gene?, put. frameshift: 202550 (-3<-1)
y4JP		-1	202406 - 203194	262							hyp. 29.5 kd protein; y4JOP: two fragments of one gene?, put. frameshift: 202550 (-3<-1)
y4JQ		+2	203729 - 206848	1039							hyp. 115.9 kd protein
y4JR		+1	206860 - 207315	151							hyp. 17.3 kd protein
y4JS		+1	207316 - 208557	413							hyp. 44.8 kd protein
y4JT		-1	208877 - 209887	336	17-283	Y4JH	239	this work	32	53	hyp. 36.4 kd protein precursor
y4KA		-3	209917 - 210885	322							hyp. 36.7 kd protein
y4KB		+1	211663 - 212088	141							hyp. 15.2 kd integral membrane protein
r2		-1	212111 - 212479	122	58-116	ORF14	104	X00493	59	76	hyp. fragment; sim. to Y4hP, Y4jD and Y4qI: additional homology to ORF14 in fr. +3/+2: 212331-212509
y4KD		-1	212750 - 214399	549							hyp. 60.4 kd protein
y4KE		-1	214412 - 215455	347							hyp. 38 kd protein; y4KEF: two fragments of one gene?, put. frameshift: 215616 (-1<-2)
y4KF		-2	215439 - 216743	434							hyp. 47.4 kd protein; y4KEF: two fragments of one gene?, put. frameshift: 215616 (-1<-2)
y4KG		-2	216855 - 217064	69							hyp. 7.7 kd protein
y4KH		-3	217105 - 217488	127							hyp. 14.1 kd protein
y4KI		-1	217670 - 218461	263							see y4bM

y4kJ	-3	218458 - 220008	516	-	-	-	-	-	see y4bL
y4kK	-1	220103 - 221041	312						hyp. 34.9 kd protein
y4kL	-2	221049 - 222041	330	101-296	ORF300	300	U23723	39	hyp. 37.6 kd AAA-family ATPase protein
y4kM	+2	222641 - 222994	117						hyp. 13.1 kd protein
y4kN	+2	223115 - 223537	140						hyp. 15.7 kd protein
y4kO	+2	223970 - 224218	82						hyp. 9.2 kd protein
y4kP	+1	224215 - 224505	96						hyp. 11 kd protein
y4kQ	-2	224898 - 225326	142						hyp. (fragmentous?) 15.3 kd protein; homology to <i>hipO</i> fragments on the complementary strand
fkl	+3	225094 - 225473					Z36940		fragments hom. to HipO
y4kR	-3	225535 - 225666	43	1-36	ORF6	347	M87280	55	hyp. 4.8 kd (fragmentous?) protein (smallest ORF predicted to be a protein); hom. to N-term. of protein in <i>crtE-crx</i> intergenic region
y4kS	-3	225751 - 226656	301	1-301	ORF8	300	U12678	93	hyp. 33.2 kd protein
y4kT	-2	226653 - 228203	516	1-516	ORF7	516	U12678	93	hyp. 55.1 kd protein
y4kU	-3	228514 - 229512	332	1-332	ORF6	332	U12678	90	prob. geranyltransferase
y4kV	-3	229666 - 231009	447	92-447	CYP117	356	U12678	89	cytochrome P-450 BJ-4 homolog
y4IA	-2	231009 - 231845	278	1-274	ORF4	275	U12678	83	short-chain type dehydrogenase/reductase
y4IB	-3	231832 - 232140	102	1-58	ORF3	94	U12678	93	put. P450-system 3Fe-3S ferredoxin
y4IC	-2	232170 - 233573	467	48-428	CYP114	382	U12678	90	cytochrome P-450 BJ-3 homolog
y4ID	-1	233666 - 234868	400	3-400	CYP112	401	U12678	92	cytochrome P-450 BJ-1 homolog
f3	-2	235704 - 235904	66	2-54	ORF8	>207	X66124	60	hyp. 7.6 kd protein fragment, homology to ORF8 fragments also upstream of f3 up to 236048
f1I	-	236796 - 237416					Z36981		homology to hupK/hupJ fragments (fr. -3/-2)
y4IF	+1	237508 - 238479	323						hyp. 36.1 kd protein
y4IG	+2	238490 - 238975	161						hyp. 17.4 kd protein
y4IH	-2	238959 - 239537	192	3-184	Fic	200	M28363	34	hyp. 22.4 kd protein; hom. to cell filamentation/division protein
y4II	-2	239541 - 239750	69						hyp. 7.3 kd protein
y4IJ	-3	240358 - 240861	167						hyp. 18.1 kd protein
f2	-	240920 - 241040					X65471		fragments of transposase (ISRm4)
y4IK	+1	241207 - 241605	132						hyp. 14.3 kd protein
y4IL	-2	241845 - 244328	827	118-816	SLR0359	1244	D63999	33	hyp. 91.8 kd protein (member of <i>E. coli</i> YegE/YhdA/YhjK/YjcC family)
f4	+1	244540 - 244851	103	19-103 28-81	TnpA F15	990 112	L14931 this work	39 94	put. truncated transposase; hom. to N-term. of TnpA (transposon Tn163); strong similarity to C-terminus of F15
y4IN	+3	244848 - 245330	160						hyp. 18.1 kd protein

y4IO		-3	247156-247938	260	11-216	AviRxxv	373	L20423	36	50	hyp. 29.1 kd protein; hom. to avirulence protein; put. frameshift according to homolog: 247230-247293 (-2<-3); end of shifted frame: 246960
f5		+1	248290 - 248628	112	59-112	F14	103	this work	94	98	hyp. protein fragment; strong similarity to part of Fl4
f6		+3	248814-249680	288	8-286	Tnp	988	M97297	27	49	put. fragmentous transposase; homologous to C-term. of transposase (Tn1546)
y4IR		+3	249696 - 251264	522							hyp. 56.8 kd protein
y4IS		+1	251407-251958	183	3-176 4-181	PaeR7IN Y4cG	195 305	S78872 this work	42 40	56 60	put. integrase/recombinase ("resolvase-type")
y4mA		+3	251955 - 252380	141							hyp. 15.8 kd protein
fm1		-	254694 - 254920								fragments hom. to xylitol-dehydrogenase
y4mB		+3	255450 - 256139	229	59-229	ORF4	212	X13583	33	53	hyp. 24.6 kd outer membrane protein precursor
y4mC		+2	256811 - 257524	237							hyp. 26.2 kd protein precursor
y4mD		-1	258065 - 258334	89							hyp. 10 kd protein
y4mE		-3	259030 - 260292	420	6-334 2-417	HipA Y4mM	440 409	M61242 this work	32 34	46 56	hyp. 45.7 kd protein
y4mF		-2	260289 - 260519	76	11-47	ORF3	90	X06090	37	70	hyp. transcriptional regulator; very low similarity to phage repressor proteins
y4mG		+3	261174 - 261395	73							hyp. 7.8 kd protein
y4mH		-2	261747 - 262640	297							hyp. 33.9 kd protein
y4mI		-2	262698 - 263672	324	11-252	RbsB	296	M13169	25	49	prob. ABC transporter periplasmic binding protein precursor (transport system Y4mJK probably transports a sugar)
y4mJ		-3	263716 - 264717	333	12-323	RbsC	321	M13169	34	55	prob. ABC transporter permease
y4mK		-2	264714 - 266207	497	8-489	RbsA	501	M13169	34	55	prob. ABC transporter ATP-binding protein
y4mL		-3	266218 - 267477	419	1-418	HI1029	425	U00079	33	58	put. permease (<i>E. coli</i> YiaN/YgiK family)
y4mM		-2	267474 - 269099	541	38-360	HI1028	328	U32729	33	54	put. permease (SBR family 7)
y4mN		-1	269096 - 270133	345	37-340	Tkt	655	U09256	36	54	hyp. transketolase family protein (fragmentous?); hom. to C-term. of transketolases
y4mO		-3	270130 - 270969	279	9-270	Tkt	655	U09256	36	52	hyp. transketolase family protein (fragmentous?); hom. to N-term. of transketolases
y4mP		-3	271000 - 271761	253	4-249	F09E10.3	255	U41749	41	60	put. short-chain type dehydrogenase/reductase
y4mQ		+1	271909 - 272805	298	1-289	PerR	297	U57080	48	65	hyp. transcriptional regulator (LysR family)
y4nA		-2	273204 - 275384	726	45-302 365-718	ORF	690	D14005	21 38	36 54	prob. peptidase; very low similarity to Y4qF and Y4sO (<25% identity)
y4nB	nodU	-3	276451 - 278127	558	1-558	NodU	558	X89965	100	100	inv. in 6-O-carbamoylation of Nod factors; similar to Y4hD
y4nC	nodS	-1	278144 - 278794	216	1-216	NodS	216	J03686	100	100	methyltransferase inv. in Nod-factor synthesis

y4nD		-3	280453 - 281349	298	-	-	-	-	-	-	see Y4iQ
y4nE		-2	281346 - 282860	504	-	-	-	-	-	-	see Y4jA
fn1		+	283238 - 283467				241	M26938			hom. to virG fragments; similar to fq3
y4nF		+3	283809 - 284501	230							hyp. 25.4 kd protein precursor; low similarity to Y4aO (<30% id.)
fn2		-	284752 - 284923					X79443			fragments hom. to ORF2 (IS-ATP-binding protein) from IS1162
y4nG		+2	285407 - 286597	396	53-365	ORF4	333	U08223	31	47	put. NAD-dep. nucleotide sugar epimerase/dehydrogenase
y4nH		+1	286594 - 286947	117	5-113	MvrC	110	M62732	30	47	hyp. 12.3 kd integral membrane protein (some similarity to ethidium bromide resistance proteins)
y4nI		+2	286964 - 287326	120							hyp. 13 kd transmembrane protein
y4nJ		+1	287335 - 288852	505	80-266 343-468	BetaA	548	U39940	29 32	44 45	hyp. GMC-type oxidoreductase
y4nK		-2	288906 - 290894	662							hyp. integral membrane protein
y4nL		-3	290914 - 291984	356	14-345	ORF6	328	U47057	26	45	put. NAD-dep. nucleotide sugar epimerase/dehydrogenase
y4nM		-3	292003 - 293553	516	226-514	NoeC	307	L18897	30	52	put. permease
y4oA		-3	294502 - 296283	593	328-494 4-590	MccB Y4qC	350 583	X57583 this work	29 30	41 50	hyp. 65.2 kd protein; homolog inv. in production of the translation inhibitor microcin C7
y4oB		+1	296572 - 296961	129							hyp. 14.7 kd protein
y4oC		+1	296965 - 297657	230							hyp. 26 kd protein
y4oD		-1	297746 - 298390	214							hyp. 23.5 kd protein
y4oE		-3	298939 - 299148	69							hyp. 7.4 kd protein
fo1		-2	299145 - 299588	147							fo1 and fo2: two fragments of one put. gene; put. frameshift: 299664 (-2<-3)
fo2		-3	299578 - 299955	125	25-109 1-123	ORF11 Y4cM	344 325	X53264 this work	37 25	63 51	homology to 5' part of ORF11; fo1 and fo2: two fragments of one putative gene; put. frameshift: 299664 (-2<-3)
fo3		+3	300015 - 300815	267	15-252	Tnp	518	L09108	40	59	fo3 and fo7: transposase-like protein interrupted by NGRIS-6
fo4		-2	300828 - 301259	143	1-143	Y4bA	694	this work	77	83	hyp. fragment; fo4/5/6: fragments of one gene similar to Y4bA/Y4pH
fo5		-1	301274 - 301684	136	1-127	Y4bA	694	this work	83	94	hyp. fragment; fo4/5/6: fragments of one gene
fo6		-2	301608 - 302900	430	1-393	Y4bA	694	this work	89	95	hyp. fragment; fo4/5/6: fragments of one gene
y4oL		-3	302890 - 303156	88	1-88	Y4bB	98	this work	63	69	hyp. 9.6 kd protein
y4oM		-1	303179 - 303628	149	1-149	Y4bC	149	this work	79	88	hyp. 16.8 kd protein
y4oN		-2	303810 - 304022	70	1-70	Y4bD	89	this work	73	84	hyp. 8.1 kd protein
fo7		+2	304118 - 304453	111	4-103	Tnp	518	L09108	40	59	fo3 and fo7: transposase-like protein interrupted by NGRIS-6

y4oP		+1	304861 - 306156	431	47-429	ul756v	469	U15180	27	42	prob. ABC transporter binding protein (Y4oPQRS: sugar-like transport system)
y4oQ		+2	306236 - 307165	309	31-301	MalF	310	U15180	35	56	prob. ABC transporter permease protein
y4oR		+2	307178 - 308011	277	12-277	MalG	296	U15180	30	52	prob. ABC transporter permease protein
y4oS		+1	308008 - 309123	371	7-369	UgpC	369	U00039	50	68	prob. ABC transporter ATP-binding protein
y4oT		-2	309132 - 309722	196	2-196	Y4pA	609	this work	28	50	hyp. 20.6 kd protein; homologous to N-terminus of Y4pA, and weakly to Y4oV
y4oU		+1	309853 - 311061	402							hyp. 43.1 kd protein precursor
y4oV		+2	311051 - 311908	285	3-280	Y4pA	609	this work	32	56	hyp. 30.2 kd protein; homologous to N-terminus of Y4pA, and weakly to Y4oT
y4oW		+1	311911 - 312561	216							hyp. 23.7 kd protein
y4oX		+3	312606 - 313688	360	36-233	MocA	317	X78503	29	44	prob. NAD-dep. oxidoreductase
y4pA		+1	313714 - 315543	609	310-596 6-290 35-237	HydG Y4oV Y4oT	441 285 196	U00006 this work this work	33 32 28	50 56 50	put. transcriptional regulator (sigma54-dep.)
y4pB	otsB	+3	316350 - 317147	265	30-260	OtsB	266	X69160	41	57	prob. trehalose-6-phosphate phosphatase
y4pC	otsA	+1	317185 - 318579	464	1-456	OtsA	474	X69160	46	66	prob. trehalose-6-phosphate synthase; similar to fq1/2 fragments homologous to ORF3; put. frameshift acc. to homologue: 319122 (3>1)
fp1		+	318915 - 319242					U08864			fragment homologous to ORF1 from IS1248 (fr. 3); similar to fs4
fp2		+	319236 - 319670					U08864			put. transcriptional regulator (MucR family); missing Zn finger motif; similar to Y4aP
y4pD		-1	319601 - 320116	171	13-140	Ros	142	M65201	50	71	identical to y4sA; hyp. 15.5 kd protein hom. to N-term. of RFRS9 25kDa protein
y4pE		-1	320606 - 321013	135	1-135		222	U18764	91	94	identical to y4sB; put. transposase; low similarity to Y4qE, Y4iB and Y4iO (<30% aa-id.)
y4pF		-2	321297 - 322460	387	50-374	Trp	334	Z48244	43	60	identical to y4sC; hyp. 21.1 kd protein
y4pG		-3	322486 - 323064	192	1-191	ORFA	197	U22323	47	64	"ORF" homologous to ORF1 of IS1162 interrupted by stop codon (323444)
fp3		+2	323189 - 323956					X79443			see y4bA
y4pH		-1	323969 - 326053	694	-	-	-	-	-	-	see y4bB
y4pI		-2	326043 - 326309	88	-	-	-	-	-	-	see y4bC
y4pJ		-3	326329 - 326778	149	-	-	-	-	-	-	see y4bD
y4pK		-1	326969 - 327238	89	-	-	-	-	-	-	fragment homologous to put. IS-ATP-binding protein
fp4		+1	327277 - 328059					L09108	48	65	

y4pL		+3	328071 - 328808	245	1-204	ORF2	231	X79443	51	63	put. insertion sequence ATP-binding protein; similarity to Y4bMY4kUY4tA, Y4uH, and weakly to Y4IQ/Y4nD/Y4sD (<30 aa-id.)
y4pM		+2	329159 - 329977	272				this work	55	73	
fp5		-	330657 - 331414					this work	61	77	
y4pN	<i>syfMI</i>	-3	332506 - 333522	338	13-324 1-338	SyrM SyrM2	326 339	M33495 this work	63 62	77 79	hyp. 30.9 kd protein put. frameshift: 331032 (2<1) probable symbiotic regulator (LysR family)
y4pO		+1	335062 - 336264	400	1-400	Tnp	400	M60971	96	98	prob. transposase (Mutator family); similarity to fe7
fq2		-2	333987 - 335003	338	1-320	OtsA	474	X69160	44	61	join fq1+fq2: hom. to trehalose-6-phosphate synthase interrupted by ISRM3-like element NGRIS-8; similarity to Y4pC (45% aa-id.)
fq1		-1	336311 - 336694	128	44-174	OtsA	474	X69160	48	67	see fq2
fq3		+	337338 - 338056					M26938			virG homologous fragments: stop at 37380; put. frameshift at 337844 (3>2); similar to fn1
y4qB		-1	339053 - 339547	164				Z54354	28	46	hyp. 18.8 kd protein
y4qC		-3	339535 - 341286	583	314-489 1-583	ORF Y4oA	401 593	this work	30	50	hyp. 63.6 kd protein
y4qD		-3	343216 - 343950	244	1-244	Y4rO	618	this work	55	74	hyp. 26.8 kd protein, similar to N-terminus of Y4rO
y4qE		+2	344114 - 345286	390	37-380	Tnp	364	X77623	38	57	prob. transposase; low similarity to Y4pF/Y4sB, Y4IB, Y4IO and Y4rJ (<30% aa-id.)
fq4		+3	345798 - 346130					M38257	34	51	fragments homologous to XerC (integrase)
y4qF		-2	346215 - 348479	754	41-725 32-736	PtrII Y4sO	707 705	D10976 this work	31 70	49 84	prob. peptidase (S9A family); high similarity to Y4sO; low similarity to Y4nA (<25% id.)
y4qG		-2	348501 - 349847	448	40-389	YgiG	454	U32722	42	62	prob. aminotransferase (class 3)
y4qH		-1	350294 - 351274	326	144-326	LasR	239	M59425	37	51	hyp. transcriptional regulator (LuxR family)
y4qI		-2	351837 - 353456	539	146-419	ORF1	322	M25805	44	63	hyp. 59.7 kd protein; similar to Y4aQ, Y4hP, Y4ID
fq5		-3	353533 - 353775								fragments fq5 and fr3 represent one put. gene similar to Y4hO and Y4IC interrupted by IS elements
y4qJ		-1	354140 - 355336	398	7-395	TnpA	388	U14952	42	60	put. transposase
y4qK		-2	355344 - 356270	308	51-293 51-308	Int Y4eF	259 251	U14952 this work	39 92	55 94	put. integrase/recombinase ("phage-type"); similar to Y4rF; low similarity to Y4rABCDEF
fq6		-2	356436 - 356636	66	1-66	Fe5 Y4rC	66 332	this work this work	79 45	94 62	put. defective integrase/recombinase ("phage-type"); 75% nt-identity: 356436-356710 and 94988-95262 [R-20]
y4rA		+1	356803 - 358032	409	17-397	ORF2	415	L34580	39	55	put. integrase/recombinase ("phage-type")
y4rB		+3	358029 - 358973	314	135-267	TnpI	284	X07651	30	51	put. integrase/recombinase ("phage-type")

y4rC		+2	358970 - 359968	332	22-294 267-332 267-332	XerC Fe5 Fq6	295 66 66	U32696 this work this work	31 41 45	50 55 62	put. integrase/recombinase ("phage-type")
y4rD		-3	360025 - 360870	281	15-277	XprB	298	M54884	25	46	put. integrase/recombinase ("phage-type")
y4rE		-2	360867 - 361799	310	50-288	YqkM	296	D84432	27	48	put. integrase/recombinase ("phage-type"); low similarity to Y4gA
y4rF		-1	361796 - 363073	425	126-414	ORF2	415	L34580	34	49	put. integrase/recombinase ("phage-type")
y4rG		-1	363287 - 363694	135	16-109	ORF1	130	U19148	32	48	hyp. 14.8 kd protein (IS866 family); low similarity to Y4jB, Y4hN
y4rH		-3	363895 - 365331	478	62-374	Bcp	598	X63470	26	44	put. ligase; hom. to biotin carboxylases
fr1		-3	366307 - 366669								85% aa-identity to part of Y4rL
fr2		-	366594 - 367402								put. frameshift: 367296 (-2<-1)
fr3		-3	367705 - 367827								hom. to N-term. of Y4hO; see fq5
y4rI		-3	368503 - 369675	390							hyp. 44 kd protein
y4rJ		+1	369697 - 370887	396	152-379	Tnp	339	M80806	28	45	put. transposase; low similarity to Y4qE (<30% aa-id.)
y4rK		-1	370976 - 371350	124							hyp. 14.5 kd protein
y4rL		-2	371454 - 371921	155	1-99 17-155	Y4zA Y4jE	295 135	this work this work	99 33	99 52	hyp. 17.7 kd protein; y4rLM: two fragments of one gene?; put. frameshift: 371972 (-2<-3); 85-99% aa-identity to parts of Y4zA and fr1
y4rM		-3	371938 - 372990	350	258-339	Y4zA	295	this work	98	98	hyp. 39.4 kd protein; see y4rL
y4rN		-2	373578 - 374795	405	35-368	P43	416	X57470	26	44	hyp. 41.6 kd integral membrane protein
y4rO		+1	375313 - 377169	618	274-596 1-244	HIN0578 Y4qD	366 244	U32742 this work	25 55	45 74	hyp. 69.3 kd protein; N-terminus: hom. to Y4qD; C-terminus: hom. to C-terminus of histidinol-1-phosphate transaminase
fr4		+	377185 - 377534					X66016			sim. to Y4rG; put. frameshift: 377376 (1>3); hom. to fragment of ORFA3 (377409 - 377540)
y4sA		-3	377842 - 378249	135	-	-	-	-	-	-	see y4pE
y4sB		-1	378533 - 379696	387	-	-	-	-	-	-	see y4pF
y4sC		-2	379722 - 380300	192	-	-	-	-	-	-	see y4pG
y4sD		-1	380933 - 381829	298	-	-	-	-	-	-	see y4iQ
y4sE		-3	381826 - 383340	504	-	-	-	-	-	-	see y4jA
fs5		-3	383593 - 384054	153	8-150	Tnp	334	Z48244	48	65	put. defective transposase; sim. to fs1

			384210 - 384493															fragments with 94-84% nt-id. to ISRM6 (<i>R. meliloti</i> ; acc. no. X95567)
y4sG	+1		384808 - 385818	336	97-325	Ddl	306	M14029	34	57								hom. to D-alanine:D-alanine ligase; probably different function
y4sH	+3		386505 - 387890	461	267-337	CapA	411	M24150	42	63								hom. to encapsulation protein A; nearly identical to Y4uA
fsI	-		388138 - 388586			Tnp		Z48244										fragments of put. transposase; put. frameshift: 388452 (-3<-2); sim. to Y4pF, Y4sB, fs5
fs2	+2		388697 - 388897			ORF1		U19148	43	62								put. transposase fragment; hom. to N-term. of ORF1; sim. to Y4jB, Y4rG, Y4hN
fs3	+		388966 - 390695			AtoC		U17902										put. transcriptional regulator fragment (put. frameshifts: 389891 (1>2); 390170 (2>3)); sim. to Y4pA, Y4oV, Y4oT)
y4sl	+2		390971 - 393241	756	325-741 1-749	McpA Y4fA	657 845	X66502 this work	41 29	60 49								prob. methyl-accepting chemotaxis protein
y4sl	-3	gapD	393202 - 394677	491	29-489	GabD	482	M88334	58	75								prob. succinate-semialdehyde dehydrogenase
y4sK	-1		394790 - 395170	126	5-122	C23G10.2	185	U39851	55	71								bel. to the YER057C/YIL051C/YJGF family; probably important cellular function
y4sL	-1		395204 - 395815	203	2-203	DadA	432	L02948	57	74								either functional dehydrogenase or non-functional fragment; hom. to small subunit of D-aminoacid dehydrogenase
y4sM	+1		395935 - 396318	127	1-127	ORF1	127	X74314	99	99								put. transcriptional regulator (AsnC/Lrp family; low homology to y4tD); missing H-T-H region
y4sN	+1		396523 - 396900	125	1-123	ORF2	>123	X74314	98	98								similar to ORFs derived from insertion elements (IS6501 family); low similarity to fu4
fs4	+		396855 - 397283	(143	8-141 1-141	ORF1 Fp2	186 145	X53945 this work	48 39	63 62								put. IS-derived protein fragment (homology to C-term. of ORF1 from IS869)
y4sO	-2		397608 - 399725	705	10-694 1-705	PprII Y4qF	706 754	D10976 this work	32 70	49 84								prob. peptidase (S9A family); low similarity to Y4nA (<25% id.)
fIt	+3		400377 - 400625	(83)	20-83	Y4IE	300	this work	64	78								fIt1 and fIt2: one put. gene encoding an amino acid ABC transporter binding protein interrupted by NGRIS-3c; see y4bm
y4IA	-3		400732 - 401523	263	-	-	-	-	-	-								see y4bl
y4IB	-2		401520 - 403070	516	-	-	-	-	-	-								see ft1
ft2	+1		403249 - 403899	(216)	5-195 2-215	ArgT Y4IE	260 300	V01368 this work	25 76	48 86								put. transcriptional regulator (AsnC/Lrp family; but low homology to y4sm)
y4ID	+1		404182 - 404691	169	11-161	HIN1362	168	U32817	38	64								

y4uH		-1	430538 - 431284	248	1-202	ORF2	231	X79443	48	63	put. insertion sequence ATP-binding protein; similarity to Y4pL, Y4bM/Y4kI/Y4tA and Y4iQ/Y4nD/Y4sD (IS21/IS1162 family)
y4uI		-3	431296 - 432840	514	1-514	Tnp	518	L09108	44	63	put. transposase; similarity to Y4bL/Y4kI/Y4tB (IS21/IS1162 family)
fu4		-	433222 - 433560			Tnp	201	X65471			put. transposase fragments (74-92% id. in 88 aa); 79% nt-identity to 5'term. of ISRM4
y4uJ	<i>fixU</i>	-1	433880 - 434110	76	1-70	FixU	70	X51963	63	80	hyp. 8.5 kd protein
y4uK	<i>nifZ</i>	-3	434107 - 434433	108	6-79	ORF2	>78	X07567	52	78	put. nitrogen fixation NifZ protein
y4uL	<i>fixN</i>	-2	434517 - 434711	64	1-64	FixN	64	M21841	79	84	prob. 4Fe-4S ferredoxin
y4uM	<i>nifB</i>	-1	434753 - 436234	493	1-493	NifB	490	M15544	72	81	involved in FeMo cofactor biosynthesis
y4uN	<i>nifA</i>	-1	436460 - 438244	594	37-594	NifA	584	U31630	62	74	positive regulator of <i>nif</i> , <i>fix</i> , and additional genes (sigma54-dep.)
y4uO	<i>fixX</i>	-2	438297 - 438590	97	2-97	FixX	98	M15546	84	89	prob. 3Fe-3S ferredoxin inv. in nitrogen fixation
y4uP	<i>fixC</i>	-1	438605 - 439912	435	1-435	FixC	435	M15546	82	89	required for nitrogenase activity
y4vA	<i>fixB</i>	-2	439923 - 441032	369	18-363	FixB	353	M15546	79	87	putatively inv. in a redox process in nitrogen fixation
y4vB	<i>fixA</i>	-2	441042 - 441899	285	1-280	FixA	292	M15546	75	90	putatively inv. in a redox process in nitrogen fixation
fv1		-1	442181 - 442252			NifS	384	X68444			put. NifS fragment (70% identity in 24 aa)
y4vC		-1	442316 - 442636	106	1-106	ORF118	118	X13691	54	72	hyp. 11 kd protein (HesB/YadR/YfhF family); homologues located upstream of <i>nifS</i>
y4vD		-2	443313 - 443879	188	5-173	HIN1693	241	U32848	46	60	put. redox enzyme (hom. to glutaredoxin-like membrane protein and peroxysomal membrane proteins)
y4vE	<i>nifQ</i>	+1	444337 - 445029	230	56-212	NifQ	180	M26323	39	56	putatively involved in Mo cofactor processing
y4vF	<i>dctA1</i>	+2	445088 - 446602	504	1-443	DctA1	456	S38912	99	99	C ₄ -dicarboxylate transport protein; nt-deletion at 446416 in comparison to sequence of acc. no. S38912 causing a frameshift (DctA1 is 48 aa longer than DctA1 in S38912)
y4vG		+1	446599 - 447843	414	13-413	CamC	415	M12546	34	50	prob. cytochromeP450
y4vH		+1	447844 - 448500	218	(32-157)	LinA	155	D90355	28	46)	hyp. 24.6 kd protein (with very weak homology to gamma-hexachlorocyclohexane-dechlorinase)
y4vI		+3	448557 - 450203	548	9-250	FabG	244	U39441	38	56	short-chain type dehydrogenase/reductase
y4vJ		+2	450341 - 451396	351	276-513	LuxA	357	M36597	30	48	put. monooxygenase; similar to Y4wF;
y4vK	<i>nifH1</i>	+1	451993 - 452883	296	1-188	NifH	296	M26961	99	99	Fe protein of nitrogenase
y4vL	<i>nifD1</i>	+1	452980 - 454494	504	199-393	NifD	>195	M26962	98	99	alpha-subunit of MoFe protein of nitrogenase
y4vM	<i>nifK1</i>	+3	454590 - 456131	513	132-195	NifK	>64	M26963	100	100	beta-subunit of MoFe protein of nitrogenase
y4vN	<i>nifE</i>	+1	456187 - 457677	496	1-469	NifE	547	X56894	62	78	involved in FeMo cofactor biosynthesis

y4vO		nifN	+1	457687 - 459096	469	1-455	NifN	441	M18272	70	81	involved in FeMo cofactor biosynthesis
y4vP		nifX	+3	459093 - 459575	160	22-156	NifX	159	X17433	52	68	nitrogen fixation protein
y4vQ			+3	459579 - 460067	162	22-162	ORF4	156	X17433	49	70	hyp. 17.7 kd protein, similar to proteins of other
						1-162	Y4xD	162	this work	61	75	nitrogen-fixing bacteria and to Y4xD
y4vR			+1	460501 - 460920	139	1-58	NifH	296	M26961	50	63	similar to N-term. of Fe protein of nitrogenase
y4vS		fdxB	+2	461228 - 461545	105	1-88	ORF5	102	M26323	52	65	prob. 4Fe-4S ferredoxin
y4wA			+1	463201 - 464739	512	86-499	PqgE	709	L43135	50	70	hyp. zinc protease (M16 family); sim. to Y4wB
y4wB			+3	464736 - 466079	447	236-438	PqgF	213	L43135	42	61	put. protease (lacks Zn-binding site; M16 family); sim. to Y4wA
y4wC			+3	466590 - 467021	143	8-132	ORF3	127	L13845	48	66	put. DNA-binding protein; high similarity to Y4aM
						1-143	Y4aM	143	this work	69	77	
y4wD			+1	467758 - 468891	377	11-370	MosC	407	U23753	29	48	permease-type protein; hom. to membrane protein from the rhizopine biosynthesis (<i>mosABC</i>) gene cluster
y4wE			+3	469311 - 470417	368	20-361	His1	356	D14440	32	53	prob. aminotransferase (class 2)
y4wF			+1	470824 - 471852	342	40-194	LuxA	354	X06758	27	54	put. monooxygenase; sim. to Y4vJ
y4wG			+2	471890 - 472435	181							
y4wH			+3	473343 - 473780	145	1-145	ORF2	145	M19352	64	76	hyp. 19.4 kd protein
y4wI			-2	473928 - 475469	513							hyp. 15.6 kd protein
y4wJ			-2	475503 - 475880	125							hyp. 59 kd protein
y4wK		nifW	-1	476519 - 476971	150	12-118	NifW	108	M86823	50	63	hyp. 13.3 kd protein
												NifW protein homolog; required for full activity of FeMo protein
y4wL		nifS	-2	477135 - 478298	387	4-387	NifS	402	M17349	58	73	prob. NifS protein (member of class-5 pyridoxal-phosphate-dep. aminotransferase family)
y4wM			-2	479145 - 481136	663	225-620	YejA	>409	U00008	38	55	put. ABC transporter binding protein (transporter or enzymatic function)
fwI			-1	481460 - 481834	124	1-116	DctA	441	M26531	55	61	hyp. truncated transporter-like protein; hom. to N-term. of DctA (see y4vF); two frameshifts acc. to homologue: 481606 (-3<-1); 481530 (-2<-3; homology stops at 481419)
y4wO			-3	481834 - 482154	106							hyp. 11 kd protein
y4wP			+2	482540 - 482947	135							hyp. 14.9 kd protein
y4xA		nifH2	+1	483871 - 484761	296	1-296	NifH	296	M26961	99	99	Fe protein of nitrogenase
y4xB		nifD2	+1	484858 - 486372	504	199-393	NifD	>195	M26962	98	99	alpha-subunit of MoFe protein of nitrogenase
y4xC		nifK2	+3	486468 - 488009	513	132-195	NifK	>64	M26963	100	100	beta-subunit of MoFe protein of nitrogenase
y4xD			+3	488262 - 488750	162	22-162	ORF4	156	X17433	47	73	hyp. 18 kd protein; similar to proteins of other nitrogen-fixing bacteria and to Y4vQ
						2-162	Y4vQ	162	this work	61	75	
y4xE			+1	488773 - 488976	67	1-64	ORF1	69	X55450	40	67	hyp. 7.6 kd protein; similar to proteins of other nitrogen-fixing bacteria

		+3	488973 - 489149	58	14-83	ExoX	98	M61751	31	52	hyp. 6.5 kd protein put. exopolysaccharide production repressor (intrgal membrane protein)
y4xF		+2	489281 - 489583	100							
y4xQ		+2	490010 - 491527	505							hyp. 55.5 kd protein
y4xG		-2	491655 - 492593	312	1-312 1-310	NodD2 NodD1	312 322	L38460 this work	99 68	99 83	transcriptional regulator (<i>LysR</i> family); high similarity to Y4aL (NodD1)
nodD2		+2	494297 - 494977	226	1-224	PmrA	222	L13395	39	58	signal transduction-type regulator
y4xI		+1	495157 - 496428	423	76-378	GPiV	426	J02451	27	46	hyp. protein hom. to proteins of the general secretion pathway (pulD family), sim. to Y4yD (NolW)
y4xJ		+1	496438 - 497004	188							hyp. 20.6 kd protein precursor
y4xK		-1	497444 - 498460	338							hyp. 37.1 kd protein
y4xL		-1	498719 - 499933	404	23-403	ORF1 (YceE)	408	X59939	22	49	permease-type protein
y4xM		-3	499930 - 501816	628	183-505	lucC	580	X76100	28	43	hyp. 71 kd protein hom. to aerobactin synthetase subunit
y4xN		-2	501816 - 502955	379							hyp. 40.9 kd protein
y4xO		-1	502952 - 503962	336	5-304	CysK	308	D26185	40	60	put. cysteine synthase
y4xP		-1	503963 - 505336	457							hyp. 49.9 kd protein; low similarity to diaminopimelate decarboxylase
y4yA		-3	503336 - 505800	154							hyp. 17.1 kd protein
y4yB		-2	505950 - 507740	596	1-596	NolX	596	L12251	98	99	nodulation protein as in <i>R. fredii</i> USDA257
y4yC	nolX	-3	508021 - 508725	234	1-234	NolW	234	L12251	99	100	nodulation protein (PulD family); sim. to Y4xJ
y4yD	nolW	+3	508881 - 509375	164	1-164	NolB	164	L12251	98	99	nodulation protein
y4yE	nolB	+3	509385 - 510254	289	1-289	NolT	289	L12251	96	97	nodulation protein precursor (YscJ homolog; M74011)
y4yF	nolT	+2	510251 - 510889	212	1-212	NolU	212	L12251	99	99	nodulation protein
y4yG	nolU	+3	510891 - 511517	208	1-60 73-208	ORF4 NolV	65 135	L12251	100 96	100 97	homologous to two (nodulation) proteins of <i>R. fredii</i> USDA257 (YscL homolog; M74011)
y4yH	nolV	+2	511514 - 512869	451	35-450 1-80 105-450	YscN HrcN	439 450	U00998 L12251	55 97	73 97	prob. ATPase involved in secretion
y4yI	hrcN	+1	512845 - 513381	178	1-178	ORF7	178	L12251	97	98	hyp. 20.4 kd protein
y4yJ	hrcQ	+1	513406 - 514482	358	171-350 1-358	YscQ HrcQ	307 382	L25667 L12251	27 96	46 98	prob. translocation protein inv. in secretion processes (FliN/MopA/SpaO family)
y4yL	hrcR	+2	514475 - 515143	222	6-216 1-222	YscR HrcR	217 249	L25667 L12251	46 99	66 99	prob. translocation protein inv. in secretion processes (FliP/MopC/SpaP family)
y4yM	hrcS	+1	515143 - 515418	91	1-66 1-91	YscS HrcS	88 92	L25667 L12251	34 98	65 100	prob. translocation protein inv. in secretion processes (FliQ/MopD/SpaQ family)
y4yN	hrcT	+3	515427 - 516245	272	28-250 1-272	YscT HrcT	261 272	L25667 L12251	31 98	52 99	prob. translocation protein inv. in secretion processes (FliR/MopE/SpaR family)

y4yO	hrcU	+2	516242 - 517279	345	5-339 1-340	YscU HrcU	354 351	L25667 L12251	30 99	50 99	prob. translocation protein inv. in secretion processes (FlhB/HrpN/YscU/SpaS family)
y4yP		+1	518077 - 518892	271	35-262	HipA	295	M19019	88	91	homolog is inducible by root-exudate and diadzein; frameshift acc. to homolog: 518855 (1>2)
fy1		+	519655 - 519995			NolJ	148	L26967			nodulation gene homologous fragments (80-100% id. in 97 aa); frameshifts acc. to homolog: 519789 (1>3); 519900 (3>2); 519965 (2>3)
y4yQ		+2	520280 - 521170	296							hyp. 31.3 kd integral membrane protein
y4yR		+2	521360 - 523453	697	17-677	LcdD	704	M96850	40	65	prob. translocation protein inv. secretion processes [Flagella/HR/Invasion proteins export pore (FHIPEP) family]
y4yS		+3	523470 - 524018	182							hyp. 20.1 kd protein
y4yA		+2	525005 - 525892	295	34-115 133-231	Y4rM Y4rL	350 155	this work this work	98 99	98 99	hyp. (fragmentous?) 32.9 kd protein; put. frameshift: 525699 (2>3); similar to Y4IE
y4yB		+1	526051 - 527121	356	60-320	Tnp	377	X67862	29	47	put. (fragmentous?) transposase (IS4 family) 526103 - 526200 higher cod. prob. in fr. 2; put. frameshift: 526200 (2>1)
fz1		+	527337 - 527902			Hdc	378	J02577			fragments homologous to histidine decarboxylases (30-45% id. in 134aa); put. frameshift (3>2) around 527478
y4yC		+3	529125 - 529910	261	65-248	AvrPph3	276	M86401	27	41	hyp. 28.3 kd protein; hom. to avirulence protein
y4yD		+3	530145 - 530294	49							hyp. 5.5 kd protein
fz4		+2	530432 - 530764	110	1-110	Y4jA	504	this work	72	85	hom. to C-terminus of Y4jA/Y4nE/Y4sE
fz2		+	530761 - 531250			ORFB	251	X67861			put. IS-ATP-binding protein fragments (32-40% id. in 137aa); put. frameshift acc. to homolog: 531062 (1>2)
y4yF	syrM2	+2	532676 - 533695	339	1-320 1-335	SyrM SyrM1	326 338	M33495 this work	69 62	81 79	prob. symbiotic regulator (LysR family)
fz3		+	534257 - 534422			ORF	338	M73488			fragments homologous to 1-aminocyclopropane-1-carboxylate deaminase (63-83% id. in 56aa); put. frameshift: 534291

a open reading frame (ORF)

b strand (-/+) or frame (-1; -2; -3; +1; +2; +3)

c number (no.)

d aminoacids (aa)

e GenBank/EMBL accession numbers

f identity (I) and similarity (S) have been calculated by the programme BESTFIT (local homology algorithm; Smith and Waterman, 1981) of the WISCONSIN SEQUENCE ANALYSIS PACKAGE (version 8.0, GCG, Madison, USA)

g abbreviations: prob. = probable; cod. prob. = coding probability; acc. = according; inv. = involved; sim. = similar; id. = identical; fr. = frame; acc. no. = accession number; nt = nucleotide; hyp. = hypothetical; put. = putative; hom. = homologous; dep. = dependent; N/C-term. = N/C-terminus

role of some ORFs like the luciferase-like ORFs (y4vJ and y4wF; see Table 3) in rhizobia is still not clear. In the 100 kb region, the duplication of a 5 kb sequence (position 451,886 to 456,157 and 483,764 to 488,035) including the genes *nifHDK* is remarkable. These genes encode the basic subunits of the nitrogenase. Furthermore, the transcriptional regulator *nodD2* is very interesting because its role seems not to be identical to a previously identified *nodD2* in a closely related strain (Appelbaum et al., 1988; data not shown). Also the *pmrA*-homologous ORF y4xI putatively plays an important role in regulating symbiotic processes because a *nod* box (binding region for the basic regulator *nodD1*; Fisher and Long, 1993) is located upstream of this ORF (position 493,962 to 494,000). Finally, the presence of ORFs (y4yI and y4yK to y4yN; see Table 3) homologous to type III secretion proteins, which have only been known previously in plant or animal/human pathogenic bacteria, shows that there only seems to be a subtle difference between symbiotic and pathogenic abilities of microorganisms.

In a second stage, the remaining 436 kb of pNGR234a were analyzed. Several ORFs and their deduced proteins were identified that belong to functional groups not previously identified in the analysis of cosmids pXB296, pXB368 and pXB110 (replication of the plasmid, conjugal transfer of the plasmid, functions in oligosaccharide biosynthesis and cleavage, functions in sugar or sugar-derivative metabolism, functions in lipid or lipid-derivative metabolism, functions in chemoperception/chemotaxis, functions in biosynthesis of cofactors, prosthetic groups and carriers, etc.).

Although further functional analyses of selected ORFs in pNGR234a still have to be performed, large-scale sequencing gives a global picture of their genomic organization and possible roles. Determination of putative functions of predicted genes by homology searches and identification of sequence motifs (promoters, *nod* boxes,

nifA activator sequences, and other regulatory elements) will aid in finding new symbiotic genes. High-fidelity sequence data covering long stretches of the genome are a prerequisite for these studies. The use of the dye terminator/thermostable sequenase shotgun approach has
5 allowed the completion of the entire ~500 kb sequence of pNGR234a and has opened up new avenues for the genetic analysis of symbiotic function.

0933964.082704
T02280.4966E660

Genetic Organization of the Whole Plasmid pNGR234a

Within the complete nucleotide sequence of pNGR234a, which comprises 536,165 bp, a total of 416 ORFs were predicted to encode proteins. An additional 67 ORF-fragments were detected that seem to be remnants of functional ORFs.

Thirty four percent (139) of the 416 potential proteins, have no obvious similarities to any known proteins. Of the remaining 277 proteins, 31 (8%) are similar to proteins for which no biochemical or phenotypic role has been assigned, 12 (3%) are similar to proteins for which limited biological data is available, and 234 (56%) are similar to proteins with a more precise biological function: enzymes (95), proteins involved in integration and recombination of insertion elements (44), transporters (32), transcriptional regulators (22), protein secretion/export (21), proteins involved in replication and control of the plasmid (12), electron transporters (6), and proteins involved in chemotaxis (2). A high proportion of enzymes was expected of a symbiotic replicon involved in nodulation (Nod-factor biosynthesis, etc.) and nitrogen fixation. As expected from the observation that NGR234 can be cured of its plasmid (Morrison *et al.*, 1983), no ORFs essential to transcription, translation or to primary metabolism were found.

A large number of protein families are present in several copies on pNGR234a. This is true even after elimination of the many proteins which are encoded in repeated IS elements, or are involved in transposition, integration or recombination. The most notable examples of highly represented protein families include: five members of the short-chain dehydrogenase/reductase family, one of which (y4vI) contains two homologous domains; five complete and one partial ABC-type transporter operons that each encode for at least one ABC-type permease and an ABC-type ATP-binding protein; four cytochrome P450's; and three members of peptidase family S9A. In total, 85 proteins belong to families that are represented more than once and which do not seem to be linked to insertion or recombination.

The majority (330, 79%) of the putative proteins are probably located in the cytoplasm of the bacterium, 62 (15%) possibly span membranes, 20 (5%) could be located in the periplasm, 3 are predicted to be lipoproteins that could associate with the outer membrane, and 2 are probably outer membrane proteins. These observations accord well with the dominance of biosynthetic proteins, as well as proteins involved in transcriptional regulation and insertion/recombination, most of which are thought to be cytoplasmic.

Although other start points cannot be excluded, replication of pNGR234a probably begins at *oriV* which is located within the intergenic sequence (*igs*) between the *repC* and *repB*-like genes y4cI and y4cJ. This locus (positions 54,417 to 54,570) encodes three proteins with 40-60% amino acid identities to RepABC of pTiB6S3 (a Ti-plasmid of *Agrobacterium tumefaciens*), pRiA4b (an Ri-plasmid of *A. rhizogenes*) and pRL8JI (a cryptic plasmid of *R. leguminosarum* bv. *leguminosarum*). Amongst replication regions, highest identities (69 to

71% at the nucleotide level) are found in the *igs*'s between *repC* and *repB* (Fig. 5). In *Agrobacterium*, these *igs*'s are the determinants which render parental plasmids incompatible. Two ORF's (position 198,500), which are homologous to pseudomonal genes involved in plasmid stability, may also play a role in replication of pNGR234a. A 12 bp portion of the origin of transfer (*oriT*) is identical to that of pTiC58 of *Agrobacterium tumefaciens* (nt 80,162 to 80,173), and highly similar to those of RSF1010 (*Escherichia coli*) and pTF1 (*Thiobacillus ferrooxidans*). This sequence corresponds to the *oriT* of plasmids containing the "Q-type nick-region" (Fig. 6).

Another 24 predicted ORFs show homologies to conjugal transfer genes of *Agrobacterium* Ti-plasmids. All are located in two large clusters between position 57,000 to 83,000. Since pNGR234a was believed to be non-transmissible (Broughton *et al.*, 1987), the fact that both the nucleotide sequence of the individual ORFs and their order is similar in *Agrobacterium* and NGR234 came as a surprise. Conjugal transfer of Ti plasmids in *A. tumefaciens* is controlled by a family of *N*-acyl-L-homoserine lactone auto-inducers (Zhang *et al.*, 1993). Similar molecules, which are able to interact with the *traR* gene product of *A. tumefaciens*, were detected in the supernatants of NGR234 cultures using the assay of Piper *et al.* (1993).

Reiterated sequences first became apparent in NGR234 during the construction of an ordered array of cosmid clones (Perret *et al.*, 1991). It is now clear that 97 kbp (18 %) of pNGR234a represents insertion- (IS) and mosaic- (MS) sequences (Fig. 7). Homology searches for known IS/MS revealed some of these, while comparison of repeated sequences within pNGR234a, as well as between the plasmid and 2,500 random chromosome sequences (V. Viprey, pers. communication) located the rest. Seventy five putative ORFs (18% of the total) and 40 fragments of ORFs were identified this way, nearly half of which (44) show homologies to integrases and transposases. Many of these IS elements are similar not only to those derived from *Rhizobium* and *Agrobacterium* species, but also to those of other, diverse Gram (-) and Gram (+) bacteria (e.g. *Bacillus*, *Escherichia*, and *Pseudomonas*). The sheer number and diversity of these IS/MS elements suggests that NGR234 has functioned as a "transposon trap". This is supported by the fact that their average G,C content (61.5 %) is 3 % higher than that of pNGR234a (58.5 %). Interestingly, many IS/MS are clustered between positions 300,000 to 390,000 (Fig. 7), while some loci are almost unaffected by insertions (*oriV*, *nod*-, *fix*- and *nif*-ORFs). Small IS/MS clusters divide the replicon into large blocks of often functionally related ORFs (e.g. blocks of *nod*-ORFs, replication and conjugal transfer ORFs, *nif*-ORFs and *fix*-ORFs). A list of all sequences with IS-element or mosaic sequence character is given in Table 4. Although transposition of these IS/MS elements has not been demonstrated, transfer of plasmids amongst rhizobia in the legume rhizosphere (Broughton *et*

Table 4: Insertion/mosaic sequences in pNGR234a

start of region	stop of region	name of region	put. ORFs/ ORF-fragments included	similarities within pNGR234a	similarities to chromosome	homologous sequences in other organisms/comments
17000	17600	ISH-10b	y4aQ	33% aa-id. to y4HP (ISH-10a)		gene products from IS866 and IS66 from <i>Ag. tumefaciens</i>
18900	19661	ISH-11b	fa2	54% aa-id. to part of y4bF (ISH-11a); 19096-19362: 91% nt-id. to ISH-11c		Tnp of IS1202 from <i>Str. pneumoniae</i>
19666	22981	NGRIS-4a	y4bABCD	identical to NGRIS-4b	many copies on the chromosome	
22985	25400	ISH-11a	fb1, y4bF	y4bF: sim. to fb1 and fa2 (ISH-11b)	partially 91% nt-id. to chromosomal sequences	Tnp of IS1202 from <i>Str. pneumoniae</i>
32463	35085	NGRIS-3a	y4bLM	identical to NGRIS-3b/c	copie(s) on the chromosome	62% nt-id. (over 2332 nt) to IS1162 of <i>Ps. fluorescens</i> (IS21/IS1162/IS408 family)
49300	50300	ISH-13a	y4cG	similar to y4IS (ISH-13b)		DNA invertase
69936	70385	ISH-4c	fd1	70233-70385: 93% nt-id. to part of NGRIS-4		ORFA of IS5376 from <i>B. stearothermophilus</i>
93322	96025	ISH-12a	fe2, y4eF, fe5, fe3	93574-94927: 90% nt-id. to ISH-12b1; 75% nt-id. to fq6 region (ISH-12b2); 95343-95558: 88% nt-id. to ISH-12b3		Tnp (fe2) and Int (Y4eF, fe3) from <i>Weizsella zoohelcum</i> -IS-element; (93322-94586: 57% nt-id. to IS292 from <i>Ag. radiobacter</i>); "phage" integrase family (Y4eF, fe5, fe3)
101939	102394	ISH-8b	fe7			84% nt-id. to ISRM5 of <i>R. meliloti</i> ; fe7: mutator family of transposases
115881	116004	MSH-14b		partially homologous to ISH-14a	72-73% nt-id. to sequences downstream from <i>chl</i> /upstream from <i>rpoN</i> on the chromosome	mosaic element
124396	124500	MSH-14a		partially homologous to ISH-14b	82% nt-id. to sequence RIME1 downstream from <i>chl</i> on the chromosome; parts of MSH-14a show 73-89% nt-id. to chromosomal sequences	mosaic element
126806	127369	ISH-12f	y4gA	low similarity to y4fE		recombinase from pAE1 of <i>Al. eutrophus</i> ("phage" integrase family)
127900	128500	ISH-12e	y4gC			
131000	131800	ISH-15	y4gE*		partially 87% nt-id. to chromosomal sequences	96% nt-id. to repetitive sequence from <i>R. fredii</i> USDA257 (acc. no. M73698)
159781	160564	ISH-16				

164600	167700	ISH-10a	y4hNOPQ		99% nt-id. of parts of y4hPQ to chromosomal sequences	different ORFs derived from IS-like sequences; partially known as acc. no. X74068 ("Region2" from pNGR234a); 164853-167086: 66% nt-id. to IS66 from <i>Ag. tumefaciens</i>
168208	169190	ISH-2c	fl1, fl2, fl3	168343-168659: 72% nt-id. to ISH-2f1/ISH-2d1 168785-169091: 73% nt-id. to ISH-2f2/ISH-2d2		168208-168383: 70 nt-id. to ISRM2011-2 (<i>R. meliloti</i>); fl2/3: IS1111A, IS1328, IS1533 family of transposases
173295	173702	ISH-8g	y4IE*	y4IE: sim. to y4tL, y4zA, and fr2		
175590	175909	ISH-11c	y4IG*	175643-175909: 91% nt-id. to ISH-11a		
185672	186507	ISH-2d	y4IO*/P* (3'-end)	185672-186075(-): 73% nt-id. to ISH-2c2(+) 186208-186507(-): 72% nt-id. to ISH-2c1(+)		Y4IO: Top of IS1328 from <i>Y. enterocolitica</i> (IS1111A, IS1328, IS1533 family)
187112	189752	NGRIS-5a	y4IQJA	identical to NGRIS-5b/c	copie(s) on the chromosome	IstA and B (Tnps) of IS1326 from <i>E. coli</i> (IS21/IS1162/IS408 family)
190000	193500	ISH-10c	y4JBCE(E*)	38/32 aa-id. of y4JCD to y4HOP (ISH10a)		different ORFs derived from IS-like sequences; partially 60% nt-id. to IS866 (<i>Ag. tumefaciens</i>); IS292 (<i>Ag. radiobacter</i>); ISR11 (<i>R. leguminosarum</i>)
193518	193634	MSH-17				76% nt-id. to repetitive sequence RMX6 from <i>Myxococcus xanthus</i> (acc. no. M60865)

199746	199938	ISH-11d	y4JM*	similarity to fb1 and y4bE (ISH-11a)	74% nt-id. to ISRI1 (<i>R. leguminosarum</i>), IS66/IS866 derivative
211165	211265	ISH-10g			74% nt-id. to IS66
212350	212580	ISH-10h	fr2	similar to y4JD (ISH-10c)	62% nt-id. (over 2352 nt) to IS1162 of <i>Ps. fluorescens</i> (IS21/IS1162/IS408 family)
217564	220186	NGRIS-3b	y4KJ	identical to NGRIS-3a/c	IS110 family
224547	224995	ISH-18a	y4kQ (3'-end)	83% nt-id. to ISH18b (427651-428102)	
240800	241040	ISH-24b	fr2		60% nt-id. to ISRI2 from <i>R. leguminosarum</i> TnpA from Tn163 (<i>R. leguminosarum</i>)
244540	244851	ISH-19a	fr4	244620-244812: 97% nt-id. to ISH-19b	
248290	248655	ISH-19b	fr5	248463-248655: 97% nt-id. to ISH-19a	
248814	249680	ISH-20	fr6		Tnp of Tn1546 (<i>Enterococcus faecium</i> ; Tn21/501/1721 family)
251407	252400	ISH-13b	y4ISmA	y4IS: similar to y4cG (ISH-13a)	y4IS: invertase; 58% nt-id. (251409-252211) to Tn501 from <i>Ps. aeruginosa</i> (acc. no. Z00027)
258551	258657	MSH-21			mosaic sequence; 82% nt-id. to sequence upstream of <i>ropA2</i> (<i>R. leguminosarum</i> ; acc. no. X80794)
280403	283043	NGRIS-5b	y4nDE	identical to NGRIS-5b/c	IS1A and B (Tnps) of IS1326 from <i>E. coli</i> (IS21/IS1162/IS408 family)
284722	284985	ISH-1b	fr2		60% nt-id. to IS1162 (<i>Ps. fluorescens</i> , IS21/IS1162/IS408 family)
300017	300819	ISH-1c	fo3		61% nt-id. IS408 (<i>Ps. cepacia</i> ; IS21/IS1162/IS408 family)
300820	304117	NGRIS-6	fo4/5/6, y4oL/M/N	77% nt-id. to NGRIS-4	
304118	304434	ISH-1d	fo7		61% nt-id. IS408 (<i>Ps. cepacia</i> ; IS21/IS1162/IS408 family)
318854	319686	NGRIS-7	fr1-2		66% nt-id. to IS1248 of <i>Pa. denitrificans</i> interrupted by NGRIS2a and 4b; fp3/4, Y4pL:
320436	328935	NGRIS-1a	fr3/4; y4pL		IS21/IS1162/IS408 family
320590	323147	NGRIS-2a	y4pEFG	identical to NGRIS-2b	partially 88-90% nt-id. to repetitive sequence RFRS9 of <i>R. fredii</i> USDA257 (IS1111A/IS1328/IS1533 family)
323961	327276	NGRIS-4b	y4pHIUK	identical to NGRIS-4a	many copies on the chromosome (disrupts all 4 copies of NGRIS-1)
335004	336301	NGRIS-8	y4pO	similar to fe7 (ISH-8b)	88% nt-id. to ISRM3 of <i>R. meliloti</i> : mutator family of transposases
342272	342419	ISH-12d		342272-342419: 87% nt-id. to ISH- 12b4	
344100	345300	ISH-2e	y4qE		Tnp (<i>Leptospira borgpetersenii</i>): IS1111A/IS1328/IS1533 family
345755	346133	ISH-12c	fr4	345755-346133: 82% nt-id. to ISH- 12b5	Int (<i>XerC</i> , <i>E. coli</i>): "phage" integrase family
351600	351735	MSH-22			80 nt-id. to sequence from pTiS4 (<i>Ag. vitis</i> , acc. no. M91609)
351826	353794	ISH-10d	y4ql, fr5	fr5: 35% aa-id. to y4hQ (ISH-10a)	67% nt-id. to ISRI1 (<i>R. leguminosarum</i> ; acc. no. L19650); IS866/66 homolog

354000	363073	ISH-12b	y4qJK, fq6, y4rABCDEF	354942-356123/356215-356383: 90/91% nt-id. to ISH-12a1; 75% nt-id. to fcs5 region (ISH-12a2); 359743-359968: 88% nt-id. to ISH-12a3; 361029-361410: 82% nt-id. to ISH-12c; 362507-362654: 87% nt-id. to ISH-12d; low similarity to y4fB and fr4 (ISH-10cf)	70% nt-id. of parts of ISH14a1 to chromosomal sequences	Tnp and Int from <i>Weekiella zoohelcum</i> -IS-element (y4qJK), different integrases (y4rAB), integrase XerC of <i>H. influenzae</i> (y4rC); y4qK, fq6, y4rABCDEF: "phage" integrase family
363287	363694	ISH-10i	y4rG	low similarity to y4fB and fr4 (ISH-10cf)		unknown protein from IS1312 (<i>Ag. tumefaciens</i>) IS866
366252	367402	ISH-8f	fr1, fr2	366252-366524: 88% nt-id. to ISH-8c; 366773-366953: 92% nt-id. to ISH-8g; 56% aa-id. of fr3 to y4rG (ISH-10a)		75% nt-id. to IS66 (<i>Ag. tumefaciens</i>)
367699	367970	ISH-10e	fr3			
368503	369675	ISH-23	y4rI		91-93% nt-id. of parts of y4rI to chromosomal sequences	
369697	370887	ISH-2f	y4rJ	370012-370328: 72% nt-id. to ISH-2c1; 370479-370785: 73% nt-id. to ISH-2c2; 371399-371671: 88% nt-id. to ISH-8f; 371474-372228: 97% nt-id. to ISH-8d; similar to y4rG (ISH-10i); identical to NGRIS-2a		y4rJ: Tnp from IS1111a of <i>Coxiella burnetii</i> (IS1111A/IS1328/IS1533 family)
371399	372990	ISH-8e	y4rL*M*			
377185	377695	ISH-10j	fr4			377321-377695: 75% nt-id. to ISRM6 (<i>R. meliloti</i>) partially 88-90% nt-id. to repetitive sequence RFRS9 of <i>R. fredii</i> USDA257 (IS1111A/IS1328/IS1533 family)
377826	380383	NGRIS-2b	y4sABC	identical to NGRIS-2a		IS1A and B (Tnps) of IS1326 from <i>E. coli</i> (IS21/IS1162/IS408 family)
380883	383523	NGRIS-5c	y4sDE	identical to NGRIS-5ab	copie(s) on the chromosome	IS21/IS1162/IS408 family
383593	384054	ISH-2g	fr5			Tnp of IS1328 of <i>V. enterocolitica</i> (IS1111A/IS1328/IS1533 family)
384210	384493	ISH-10k				fragments with 94-84% nt-id. to ISRM6 (<i>R. meliloti</i>)
388100	388600	ISH-2h	fr1			different Tnps (IS1111A/IS1328/IS1533 family)
388601	388900	ISH-10l	fr2			ORF from IS1312 of <i>Ag. tumefaciens</i> (IS66/866 family)
396445	397301	NGRIS-9	y4sN and fr4		91-99% nt-id. of NGRIS9-parts to chromosomal sequences	different ORFs derived from IS elements; partially known from acc. no. X74314
400626	403248	NGRIS-3c	y4rAB	identical to NGRIS-3ab	copie(s) on the chromosome	62% nt-id. (over 2352 nt) to IS1162 of <i>Ps. fluorescens</i> (IS21/IS1162/IS408 family)
426525	428102	ISH-18b	y4uE*	427651-428102: 83% nt-id. to ISH-18a	77-96% nt-id. of ISH-18b-parts to chromosomal sequences	Tnp of mini-circle DNA from <i>Sir. coelicolor</i> (IS110 family)
429860	430007	ISH-8c	fr3			85% nt-id. to ISRM5 (<i>R. meliloti</i>)
430568	432851	ISH-1c	y4uHI			60% nt-id. to IS408/IS1162 (<i>Ps. cepacia/Ps. fluorescens</i>)
433222	433560	ISH-24a	fr4	low similarity to y4sN (NGRIS-9)		79% nt-id. to ISRM4 (<i>R. meliloti</i>)/ISR12-like fragments with 83-69% nt-id. to IS866 (<i>Ag. tumefaciens</i>)
462554	463053	ISH-10f				524946-525580: 61% nt-id. to ISRM5 (<i>R. meliloti</i>)
524946	525892	ISH-8d	y4zA	525095-525849: 97% nt-id. to ISH-8c		Tnp of IS3376 from <i>B. stearothermophilus</i> (IS4 family of transposases)
526051	527121	ISH-25	y4zB*			fr4/2: IS21/IS1162/IS408 family
530364	531249	ISH-1f	fr4, fr2	79% nt-id. to part of NGRIS-5		

Abbreviations: Tnp = transposase; Int = integrase; nt-id. = nucleotide-identity; aa-id. = aminoacid identity
IS elements with precisely defined borders are designated as NGRIS/NGRIS-1 to 9. Other sequences which show homologies to known mosaic or IS-like sequences (mosaic/insertion sequence homologs) are named MSH and ISH, respectively.

al., 1987) and to other non-symbiotic bacteria in fields (Sullivan *et al.*, 1995) suggests that lateral transfer of genetic information has helped shape symbiotic potential.

Carbohydrates are constituents of the rhizobial cell wall as well as morphogens called Nod-factors (short tri- to penta-mers of *N*-acetyl-D-glucosamine, substituted at the non-reducing terminus with C16 to C18 saturated or partially unsaturated fatty acids). Elements of the biosynthetic pathways leading to cell walls or to lipo-chito-oligosaccharides (Nod-factors) are common. Most differences are found in the later stages of the pathways that lead to specific cell-wall components or to Nod-factors.

As befits a symbiotic replicon, only 13 ORF's with homology to polysaccharide synthesis genes (house-keeping genes *sensu stricto*) are located on the plasmid (Table 3). Sequences homologous to *exoB*, *exoF*, *exoK*, *exoL*, *exoP*, *exoU*, and *exoX* (X. Perret and V. Vipréy, unpublished), and *exoY* (Gray *et al.*, 1990) are clearly located on the chromosome. Although loci with weak homologies to *nod*-box::*psiB* of *R. leguminosarum*, and *exoX* of *R. meliloti* exist on the plasmid (*y4iR*, and *y4xQ* respectively), these are regulatory rather than structural genes, suggesting that almost all cell wall polysaccharide synthesis ORFs are chromosomally located.

Except for *nodPQ* and *nodE*, at least one copy of all the regulatory and structural ORFs involved in Nod-factor biosynthesis seem to be located on the plasmid. The activity of most nodulation genes is modulated by four transcriptional regulators of the *lysR* family. These are *nodD1* (*y4aL*), *syrM1* (*y4pN*), *nodD2* (*y4xH*), and *syrM2* (*y4zF*). *NodC*, which is an *N*-acetylglucosaminyltransferase, the first committed enzyme in the Nod-factor biosynthetic pathway, is part of an operon which includes *nodABCIJnolOnoeIE* (*y4hI* to *y4hB*, Table 3). Together, these genes, which form the *hsnIII* locus, are responsible for the synthesis of the core Nod-factor molecule, and the adjunction of 3- (or 4)-*O*-carbamoyl, 2-*O*-methyl, and 4-*O*-sulfate groups (Hanin *et al.*, unpublished). *nodZ* (*y4aH*), which encodes a fucosyltransferase, is part of the *hsnI* locus, which includes *noeJ* (*y4aJ*), *noeK* (*y4aI*), *noeL* (*y4aG*), *nolK* (*y4aF*), all of which are involved in the fucosylation of NodNGR factors (Fellay *et al.*, 1995a). Wild-type NodNGR factors are also *N*-methylated and 6-*O*-carbamoylated, adjuncts which are added by the transferases encoded by *nodS* and *nodU* respectively [*y4nC* and *y4nB*; *hsnII* (Lewin *et al.*, 1990)]. Possibly the only other enzyme which may be directly involved in Nod-factor biosynthesis is that encoded by *nolL* (*y4eH*, Table 3). As the 2-*O*-methylfucose residue of NGR234 Nod-factors is either 3-*O*-acetylated, or 4-*O*-sulphated, an acetyltransferase is obviously required. Since *NolL* shows only limited homology to acetyltransferases, experimental proof of the transferase activity will be required however.

In contrast to *R. leguminosarum* and *R. meliloti* harbouring pNGR234a, *A. tumefaciens*(pNGR234a) transconjugants are incapable of nitrogen fixation (Broughton *et al.*, 1984), suggesting that some essential *fix*-ORFs are also carried by the chromosome.

Nevertheless, more than 40 *nif*- and *fix*-ORFs are plasmid borne. Included amongst these are *nifA* (y4uN) which encodes for a sigma-54 dependent regulator. Mutation of *rpoN* (which encodes sigma 54) causes a Fix⁻ phenotype on NGR234 hosts (van Slooden *et al.*, 1990). Similarly, mutation of *fixF* (y4gN) disrupts synthesis of a rhamnose-rich extra-cellular polysaccharide, and results in a Fix⁻ phenotype on *Vigna unguiculata*, the reference host for NGR234 (unpublished). In fact, loci adjacent to *fixF* are probably responsible for the synthesis of dTDP-rhamnose from glucose-1-phosphate. Enzymes involved in this biosynthetic pathway include glucose-1-phosphate thymidyltransferase (y4gH), dTDP-glucose-4,6-dehydratase (y4gF), dTDP-4-dehydrorhamnose-3,5-epimerase (y4gL), and dTDP-4-dehydrorhamnose reductase (y4gG). Rhamnose-rich lipopolysaccharides (LPS) seem to be necessary for complete bacteroid development and nitrogen fixation (Krishnan *et al.*, 1995). Perhaps the enzyme encoded by y4gI is needed for the synthesis of the rhamnose rich LPS's from dTDP-rhamnose.

Although not directly involved in the fixation process, mutation of the plasmid borne copy of *dctA* (= *dctA1*, y4vF) also impairs nitrogen fixation (van Slooden *et al.*, 1992). Other *nif*- and *fix*-ORFs are involved in elaboration of the electron-transfer complex (*fixAB*), in various cofactors required for nitrogen fixation (e.g. *fixC*, *nifB*, *nifE*, *nifN*, etc.), and in the synthesis of ferredoxins (*fdxB*, *fdxN*, *fixX*). Finally, those ORFs involved in the synthesis of the nitrogenase complex are also present. Amongst these are two functional copies of the *nifKDH* ORFs (y4vM to y4vK and y4xC to y4xA) (Badenoch-Jones *et al.*, 1989). Additionally, 17 new ORFs located within the nitrogen fixation cluster (see Fig. 7; ORFs y4vC to y4vJ with the exception of *dctA1*, y4wA to y4wG, y4wI, y4wJ and y4xQ) are co-transcribed together with the ORFs homologous to known *nif* and *fix* genes. It thus seems likely that most ORFs necessary for bacteroid development and synthesis of the nitrogen-fixing complex, are carried by pNGR234a.

Two types of regulatory elements which frequently occur in pNGR234a are the NodD- and NifA/sigma-54-dependent promoters. NodD-dependent promoter-like sequences known as *nod* boxes have been identified by homology search within intergenic regions, using the following consensus sequence: 5'-YATCCAYNNYRYRGATGNNNNYNATCNAAACAATCRATTTTACCAATCY-3' [12 mismatches allowed (van Rhijn and Vanderleyden, 1993); Y=C or T, R=A or G, N=A,C,G or T]. Putative NifA-dependent promoters (Fischer, 1994) have been predicted by screening for the NifA activator sequence (5'-TGT-N₁₀-ACA-3') together with the sigma-54 promoter consensus sequence (5'-TGGCAC-N₅-TTGCA/T-3' with GG and GC as the most conserved doublets; 3 mismatches allowed) separated by 60 to 150 nucleotides. The identified conserved promoter-like sequences in pNGR234a are listed in Tables 5 and 6.

Tab.5. *nod* box-like sequences in pNGR234a

<i>nod</i> box	position in pNGR234a	orien- tation	number of mismatches to the consensus sequence	distance to the following ORF	name of the following ORF
1	4514 - 4562	-	11	504	(fa1)
2	8481 - 8529	-	8	87	<i>nodZ</i>
3	12322 - 12370	-	7	-	?#
4	97470 - 97518	-	6	277	<i>nolL</i>
5	129615 - 129663	+	10	1358	<i>y4gE</i>
6	141088 - 141136	+	8	890	<i>fixF</i>
7	150280 - 150327	-	11	202	<i>noeE</i>
8	158820 - 158868	-	4	235	<i>nodA</i>
9	161891 - 161939	+	11	1103	<i>y4hM</i>
10	169833 - 169881	-	7	117	<i>y4iR</i>
11	278947 - 278995	-	7	153	<i>nodS</i>
12	279821 - 279869	+	7	-	?#
13	443101 - 443149	-	10	465	<i>y4vC</i>
14	473059 - 473107	+	9	236	<i>y4wH</i>
15 ^o	481253 - 481301	-	16	117	<i>y4wM</i>
16	493961 - 494009	+	6	288	<i>y4xI</i>
17	532039 - 532087	+	5	589	<i>syrM2</i>
18	256434 - 256482	+	12	329	<i>y4mC</i>
19	469151 - 469199	+	12	112	<i>y4wE</i>

^o The majority of the mismatches is located in the 3'-terminal part of the sequence.

No predicted ORF can be found downstream of the putative *nod* box.

Tab.6. Putative NifA-dependent promoters in pNGR234a

Nr.	NifA-dep. UAS*: position	sigma-54 promoter (-12/-24 region#): position	orien- tation	distance to the following ORF (nt)	name of the following ORF
1	90812 - 90827	90910 - 90924	+	127	y4eD
2	162727 - 162742	162788 - 162802	+	240	y4hM
3	235036 - 235051	234934 - 234948	-	66	y4lD
4	255021 - 255036	255130 - 255144	+	306	y4mB
5	285265 - 285280	285343 - 285357	+	50	y4nG
6	436363 - 436378	436275 - 436289	-	41	nifB
7	442046 - 442061	441955 - 441969	-	56	fixA
8	442735 - 442750	442676 - 442690	-	40	y4vC
9	444109 - 444124	443983 - 443997	-	104	y4vD
10	444137 - 444152	444241 - 444299°	+	38°	nifQ
11	451782 - 451799	451891 - 451905	+	88	nifH1
12	460319 - 460334	460424 - 460438	+	63	y4vR
13	463063 - 463078	463139 - 463153	+	48	y4wA
14	478839 - 478854	478761 - 478775	-	463	nifS
15°	483663 - 483678	483769 - 483783	+	88	nifH2

* "Upstream Activator Sequence": NifA-binding site located 80 to 150 nt upstream of the transcription start point (5'-TGT-N₁₀-ACA-3').

sequence corresponding to the consensus sequence of conserved sigma-54-promoters 12 nt upstream of the transcription start point: 5'-TGGCAC-N₅-TTGC-3' (2 mismatches allowed).

° 3 possibilities for a promoter (in two cases only corresponding to the minimal consens: 5'-GG-N₁₀-GC-3')

EXAMPLESExample 1

5

GENERAL METHODS**Bacteria and Plasmids**

10 *Escherichia coli* was grown on SOC, in TB or in two-
fold YT medium (Sambrook et al., 1989). The cosmid clones
pXB296 and pXB110 (Perret et al., 1991) were raised in
E. coli strain 1046 (Cami and Kourilsky, 1978). Subclones
in M13mp18 vectors (Yanisch-Perron et al., 1985) were grown
15 in *E. coli* strain DH5 α F'IQ (Hanahan, 1983).

Construction of Cosmid Libraries

20 Cosmid DNA was prepared by standard alkaline lysis
procedures followed by purification in CsCl gradients
(Radloff et al., 1967). DNA fragments sheared by sonication
of 10 μ g of cosmid DNA were treated for 10 min at 30°C with
30 units of mung bean nuclease (New England Biolabs,
Beverly, MA, USA), extracted with phenol/chloroform (1:1),
25 and precipitated with ethanol. DNA fragments, ranging in
size from 1 to 1.4 kbp, were purified from agarose gels
using Geneclean II (Bio101, Vista, CA, USA) and ligated into
*Sma*I-digested M13pm18. Electroporation of aliquots of the
ligation reaction into competent *E. coli* DH5 α F'IQ was
30 performed according to standard protocols (Dower et al.,
1988; Sambrook et al., 1989).

M13 Template Preparation

35 Fresh 1 ml *E. coli* cultures in twofold YT held in 96-
deep-well microtiter plates (Beckman Instruments, Fullerton,
CA, USA) were infected with recombinant phages from white
plaques grown on plates containing X-gal (5-bromo-4-chloro-

FO2280-4966E560

indoyl- β -D-galactoside) and IPTG (isopropyl- β -thiogalactopyranoside). Rapid preparation of ~0.5 μ g of single-stranded M13 template DNA was carried out as follows: 190 μ l portions of the phage cultures grown for 6 hr at 37°C were transferred into 96-well microtiter plates. Lysis of the phages was obtained by adding 10 μ l of 15% (w/v) SDS followed by 5 min incubation at 80°C. Template DNA was trapped using 10 μ l (1 mg) of paramagnetic beads (Streptavidin MagneSphere Paramagnetic Particles Plus M13 Oligo, Promega, Madison, WI, USA) and 50 μ l of hybridization solution [2.5 M NaCl, 20% (w/v) polyethylene glycol (PEG-8000)] during an annealing step of 20 min at 45°C. Beads were pelleted by placing microtiter plates on appropriate magnets and washing three times with 100 μ l of 0.1-fold SSC. The DNA was recovered in 20 μ l of water by a denaturation step of 3 min at 80°C. When required, larger amounts of single-stranded recombinant DNA (>10 μ g) were purified using QIAprep 8 M13 Purification Kits (Qiagen, Hilden, Germany) from 3 ml of supernatant of phage cultures grown for 6 hr at 37°C.

Sequencing

Two sequencing methods were used: dye terminator and dye primer cycle sequencing, each in combination with AmpliTaq DNA polymerase (Perkin-Elmer) and Thermo Sequenase (Amersham). All reactions, including ethanol precipitation, were performed in microtiter plates. Reagents were pipetted using 12-channel pipettes. Where necessary, sequencing reaction mixtures, including enzymes, were pipetted into the plates in advance and held at -20°C until needed.

Dye Terminator Cycle Sequencing

For dye terminator/AmpliTaq DNA polymerase sequencing, 0.5 μ g of template DNA, and the PRISM Ready Reaction DyeDeoxy Terminator Cycle Sequencing Kit (Perkin-Elmer) were used. Cycle sequencing was performed in microtiter plates

using 25 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 4 min at 60°C). Prior to loading the amplified products on electrophoresis gels, unreacted dye terminators were removed using Sephadex columns scaled down to microtiter plates
5 (Rosenthal and Charnock-Jones, 1993).

Dye terminator/Thermo Sequenase sequencing was performed using the same experimental conditions except that the reaction mix contained 16.25 mM Tris-HCl (pH 9.5),
10 4.0 mM MgCl₂, 0.02% (v/v) NP-40, 0.02% (v/v) Tween 20, 42 μM 2-mercaptoethanol, 100 μM dATP/dCTP/dTTP, 300 μM dITP, 0.017 μM A/0.137 μM C/0.009 μM G/0.183 μM T from Taq Dye Terminators (Perkin-Elmer; no. A5F034), 0.67 μM primer, 0.2 - 0.5 μg of template DNA, and 10 units of Thermo
15 Sequenase (Amersham) in a 30 μl reaction volume. Unincorporated dye terminators were removed from reaction mixtures by precipitation with ethanol.

Dye Primer Cycle Sequencing

20

Dye primer/AmpliTaQ DNA polymerase sequencing reactions were performed according to the instructions accompanying the Taq Dye Primer, 21M13 Kit (Perkin-Elmer). Cycle sequencing was carried out on 0.5 μg of template DNA
25 with 19 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 90 sec at 72°C) followed by six cycles, each consisting of 95°C for 30 sec and 72°C for 2.5 min. Prior to electrophoresis, the four base-specific reactions were pooled and precipitated with ethanol.

30

Identical PCR conditions and the Thermo Sequenase Fluorescent Labelled Primer Cycle Sequencing Kit (Amersham) were used for dye primer/Thermo Sequenase sequencing reactions.

35

Sequence Acquisition and Analysis

Gel electrophoresis and automatic data collection were

09939964-082701

performed with ABI 373A DNA sequencers (Perkin-Elmer). After removing cosmid vector and M13mp18 sequences from the shotgun sequence data, the data were assembled using the program XGAP (Dear and Staden, 1991) and edited against the fluorescent traces. To close remaining gaps, to make single-stranded regions double-stranded, and to clarify ambiguities, additional cycle sequencing reactions with selected shotgun templates were carried out using either custom-made primers (primer-walks) or universal primer.

10

The complete double-stranded DNA sequence of cosmid pXB296 was analyzed using programs from the Wisconsin Sequence Analysis Package (version 8, Genetics Computer Group, Madison, WI, USA). Homology searches were performed with BLAST (version 1.4; Altschul et al., 1990) and FASTA (version 2.0; Pearson and Lipman, 1988). Several nucleotide and protein databases were screened (GenBank/Genpept, SwissProt, EMBL, and PIR). Identities and similarities between homologous amino acid sequences were calculated with the alignment program BESTFIT (Smith and Waterman, 1981).

Example 2

25 Comparison of Fluorescent Traces Created by Different Cycle Sequencing Methods

When using a thermostable sequenase [Thermo Sequenase (Amersham)], the concentrations of dye terminators (Perkin-Elmer) can be reduced by 20- to 250-fold in comparison to the concentrations needed for *Taq* DNA polymerase without compromising the quality of the sequencing results (Table 7).

35 To compare the dye terminator and dye primer cycle sequencing procedures, representative templates derived from the pXB296 library were sequenced by both methods, each performed with Thermo Sequenase and *Taq* DNA polymerase

T.D. 230" 4966E660

Table 7. Concentrations (in μM) of dye terminators in each cycle sequencing reaction with two different thermostable DNA polymerases

Dye terminator	AmpliTaq DNA polymerase	Thermo Sequenase DNA polymerase	Dilution factor for dye terminators ^a
A Taq	0.751	0.017	40
C Taq	22.500	0.137	160
G Taq	0.200	0.009	20
T Taq	45.000	0.183	250

^aThermo Sequenase vs. AmpliTaq.

(Figure 1). In general, dye terminator traces do not contain the many compressions (on average, one compression every 50 bases in single reads) that are common with dye primers if mixes do not contain nucleotide analogues like deoxyinosine or 7-deaza-deoxyguanosine triphosphates or if sequencers are used without active heating systems. In addition, dye terminator traces obtained with Thermo Sequenase show more uniform signal intensities over those obtained with *Taq* DNA polymerase, thus resulting in a reduced number of weak and missing peaks (e.g. a weak G-signal following an A-signal in Thermo Sequenase traces or a weak C-signal following a G-signal in *Taq* DNA polymerase traces). Using ABI 373A sequencers, errors in automatic base-calling of Thermo Sequenase/dye terminator scans only arise after 300 - 350 bases. The average number of resolved bases in dye primer gels (378 bases) is 46 bases longer than in those produced with dye terminators (332 bases). Furthermore, in Thermo Sequenase/dye primer sequences the peaks are very regular and the number of stops and missing bases decreases in comparison to *Taq* DNA polymerase/dye primer electropherograms. The number of compressions, however, is not significantly reduced.

25 Example 3

Shotgun Sequencing of Entire Cosmids Using Dye Terminators or Dye Primers

30 To compare the efficiency of both methods, cosmid pXB296 of pNGR234a was shotgun sequenced using a combination of dye terminators and thermostable sequenase (Thermo Sequenase), whereas another cosmid, pXB110, was sequenced using a combination of dye primers and *Taq* DNA polymerase
35 (Table 1). Over 93% (736 clones) of 786 dye terminator reads of pXB296 were accepted by XGAP with a maximal alignment mismatch of 4%. By increasing this level to 25%, so that most of the remaining data could be included in the

assembly, 775 reads led to three 6 to 10 kbp stretches of contiguous sequence (contigs), two of which were joined after editing. To close the last gap and to complete single-stranded regions with data derived from the opposite strand, only 32 additional dye terminator reads using custom-made primers were required. It took <1 week to assemble and finalize the 34,010 bp DNA sequence of pXB296 (EMBL accession no. Z68203; eight-fold redundancy; GC content, 58.5 mol%).

10

In contrast, only 308 (34%) of 899 shotgun reads obtained by Tag DNA polymerase/dye primer cycle sequencing of pXB110 were included in the first assembly (4% alignment mismatch). At the 25% alignment mismatch level, 879 reads were assembled, leading to 25 short contigs (1 - 2 kbp). These contigs had to be edited extensively in order to join most of them. "Primer walks", covering gaps and complementing single-stranded regions, were not sufficient to clarify all the remaining ambiguities in the assembled sequence. Every 100 - 150 bp, a compression in one strand could not be resolved by sequence data from the complementary strand. Therefore, it was necessary to resequence clones using dye terminators and universal primer. In total, 191 additional dye terminator reads had to be created. As a result, assembling and finalizing the 34,573 bp sequence of pXB110 (10.5-fold redundancy; GC content, 58.3 mol%) took much more time than pXB296 did.

30 Example 4

Analysis of Cosmid pXB296

Putative ORFs were located on the 34,010 bp sequence of pXB296 using the programs TESTCODE (Fickett, 1982) and CODONPREFERENCE (Gribkov et al., 1984), the latter in combination with a codon frequency table based on previously sequenced genes of *Rhizobium* sp. NGR234 (as well as the

closely related *R. fredii*). All 28 ORFs and their deduced amino acid sequences exhibited significant homologies to known genes and/or proteins. The positions of the ORFs along pXB296, as well as the best homologues, are displayed in Table 2 and Figure 2. Ribosomal binding site-like sequences (Shine and Dalgarno, 1974) precede each putative ORF except for ORF9 (position 11,214 - 12,455). If one disregards the homology to known glutamate dehydrogenases in the first 32 amino acids deduced from this ORF, a downstream alternative start codon (position 11,220) preceded by a Shine-Dalgarno sequence can be identified. Most of the ORFs are organised in five clusters (ORFs with only short intergenic spaces or overlaps between them). Cluster I, containing ORF1 to ORF5, encodes proteins homologous to trans-membrane and membrane-associated oligopeptide permease proteins and to a *Bacillus anthracis* encapsulation protein. Cluster II, includes ORF6 and ORF7, which are homologous to aminotransferase and (semi)aldehyde dehydrogenase genes. Homologies to transposase genes [ORF8; cluster III (ORF10 and ORF11)] and to various *nif* and *fix* genes [cluster IV (ORF12 to ORF20); ORF23, part of cluster V] are also reported.

Presumed promoter and stem-loop sequences that might represent ρ -independent terminator-like structures (Platt, 1986) are shown in Figure 2. Significant σ^{54} -dependent promoter consensus sequences (5'-TGGCACG-N₄-TTGC-3'; Morett and Buck, 1989), as well as *nifA* upstream activator sequences (5'-TGT-N₁₀-ACA-3'; Morett and Buck, 1988), are found upstream of the *nifB* homologue ORF15, the *fixA* homologue ORF20, ORF21, ORF22, and ORF23. ORF23 is part of cluster V in pXB296, which includes the *dctA* gene of *Rhizobium* sp. NGR234 (van Slooten et al., 1992). Surprisingly, the published *dctA* sequence shows important discrepancies. Therefore, a fragment encompassing this locus was amplified by PCR using NGR234 genomic DNA as template. By sequencing this fragment, the cosmid sequence of the present invention was confirmed.

Example 5**Analysis of the Complete Plasmid pNGR234a**

5

Using the thermostable sequenase/dye terminator cycle sequencing method herein described, 20 overlapping cosmids (including pXB296) of the symbiotic plasmid pNGR234a of *Rhizobium* sp. NGR234 were sequenced, together with two PCR products and a subcloned DNA fragment derived from cosmid pXB564 that cover two remaining gaps (position 276,448 - 277,944 and position 480,607 - 483,991). The map of the sequenced cosmids is shown in Figure 4. The entire assembled 536 kb sequence of pNGR234a is given in Figure 3 (deposited in EMBL/GenBank under accession no. U00090).

The analysis of the complete nucleotide sequence revealed few regions of 98 - 100% identity to already published sequences in public databases. These sequences are listed in Table 8. These sequences had been derived either from *Rhizobium* sp. NGR234, derivatives of it or closely related strains of it. Therefore, the ORFs and their deduced proteins, 98 - 100% homologous to *nifH*, *nodA*, *nodB*, *nodC*, *nodD1*, *nodS*, *nodU*, *nolX*, *nolW*, *nolB*, *nolU* and "ORF1", represent already known genes/proteins (Table 8 and References). Some other ORFs and their deduced proteins, nearly identical to public database entries, were either only partially known before the disclosure of the present invention or exhibited significant differences, for instance, *dctA*, host-inducible gene A, *nifD*, *nifK*, *nodD2*, *nolT*, *nolX*, *nolV*, "ORF140", "ORF91", "RSRS9 25 kDa-protein gene" (Table 8 and References).

As a first step, approximately 100 kb of pNGR234a was analyzed between position 417,796 to 517,279 using the programs TESTCODE (Fickett, 1982) and CODONPREFERENCE (Gribskov et al., 1984). In this initial ~100 kb of sequence, 76 ORFs were found and ascribed putative functions

Table 8.
All ORFs that show 98-100% identity in the nucleotide sequence to ORFs located in pNGR234a and that have already been published in databases:

ORF	organism	EMBL/GeneBank accession no.	+	-	claimed in the patent application/ not claimed in the patent application
<i>dctA</i>	<i>Rhizobium</i> sp. NGR234	S38912	+		sequencing mistakes in the database entry: the real <i>dctA</i> in pNGR234a is 144 bases longer (see table 4)
host inducible geneA	<i>Rhizobium fredii</i> USDA 201#	M19019 RFIND	+		significant difference in pNGR234a (frameshift; see table 4)
<i>nifH</i>	<i>Rhizobium</i> sp. ANU 240*	M26961 RHMNIFKDH3	-		
<i>nifD</i> (partially)	<i>Rhizobium</i> sp. ANU 240*	M26961 RHMNIFKDH2	+		only part of <i>nifD</i> is in the public database
<i>nifK</i> (partially)	<i>Rhizobium</i> sp. ANU 240*	M26961 RHMNIFKDH1	+		only part of <i>nifK</i> is in the public database
<i>nodABC</i>	<i>Rhizobium fredii</i> USDA 257#	M73362 RSNOD2	-		
<i>nodD1</i>	<i>Rhizobium</i> sp. mpik 3030*	Y00059 RSNODD1	-		
<i>nodD2</i>	<i>Rhizobium japonicum</i> USDA 191#	M18972 RHMNODD2M	+		significantly different function of <i>NodD2</i> in NGR234 than in USDA 191 (despite of 98% identity °)
<i>nodS</i>	<i>Rhizobium</i> sp. NGR234	J03686 NGRNODSU	-		
<i>nodU</i> (partially)	<i>Rhizobium</i> sp. NGR234	J03686 NGRNODSU	-		
<i>nodU</i> (full)	<i>Rhizobium</i> sp.*	X89965 RSNODUGEN			
<i>noXWBTUV</i>	<i>Rhizobium fredii</i> USDA 257#	L12251 RHMNOLBTU	-	+	<i>noXWB</i> , <i>noIU</i> NoIT: 97% identical (amino acid sequence level) NoIX, NoIV+ORF4 of pNGR234a show significant differences to USDA257 (see table 4)
ORF1; ORF2 (partially)	<i>Rhizobium</i> sp. NGR234	X74314 RSORF	-		
ORF140 nodulation gene;	<i>Rhizobium</i> sp. NGR234	X74068 RSPLAS	+		database entry includes sequencing mistakes causing frameshifts
ORF91 (partially)			+		repetitive element in pNGR234a showing insertions, deletions of nucleotides in comparison to the database entry
RFRS9 25kDa protein gene*	<i>Rhizobium fredii</i> USDA 257#	U18764 RFU18764	+		

*strains representing derivatives of NGR234: *Rhizobium* sp. ANU 240, *Rhizobium* sp. mpik 3030, *Rhizobium* sp.

#strains closely related to NGR234: *Rhizobium fredii* USDA 257, *Rhizobium japonicum* USDA 191, *Rhizobium fredii* USDA 201.

°identity in nucleotide sequence as well as amino acid sequence

(= ORFs y4tQ to y4yO (excluding ORFs y4uD, y4uG, y4wG, y4wO, y4wP, y4xF, y4xQ, y4xG and y4yB and excluding ORF-fragments fu1, fu2, fu3, fu4, fv1 and fw1); see Table 3). It should be noted that since the sequence of cosmid pXB296 forms part of this 100 kb region, all of the ORFs identified in Table 2 (except "ORF1") are reproduced (albeit with minor, but definitive, revisions) in Table 3. Most of the 76 ORFs and their deduced proteins showed homologies to public database entries that could help identify their putative functions. Only ORFs y4vK and y4xA (duplicated *nifH*) as well as y4yD, y4yE and y4yG (*nolW*, *nolB* and *nolU*) were identical to database entries (98 - 100% homology). In the case of 7 ORFs and their deduced proteins, no homologous sequences in public databases have been found.

As a second step, the remaining 436 kb of pNGR234a were analyzed using the methods noted above. The results of this analysis are discussed in Example 6.

Example 6

Genetic Organization of the Complete Plasmid pNGR234a

In order to confirm and to improve the identification of probable coding regions in pNGR234a, the program GeneMark was used which is based on matrices developed for related organisms of *Rhizobium* sp. NGR234 (*R. leguminosarum* and *R. meliloti* (Borodovsky et al., 1994)). The use of this program currently represents the most frequently applied method to distinguish coding and non-coding regions in newly sequences DNA of prokaryotes. Further analysis of the putative ORF products was carried out using methods to detect signal sequences, transmembrane segments and various other domains (PROSITE database search (Bairoch et al., 1995); PSORT program (Nakai et al., 1991)).

In total, 416 ORFs were predicted to encode putative

proteins (Freiberg et al., 1997). Additionally, 67 fragments were detected that seemed to be remnants of functional ORFs. Some of these were disrupted by insertion of mobile elements. All identified functional ORFs and 5 fragments of former functional ORFs are listed in Table 3.

Within the initial ~100 kb region (position 417,796 to 517,279) first analyzed in this study, 9 ORFs (y4uD, y4uG, y4wG, y4wO, y4wP, y4xF, y4xQ, y4xG and y4yB) and 6 10 ORF-fragments (fu1, fu2, fu3, fu4, fv1 and fw1) were predicted in addition to the 76 ORFs (y4tQ to y4yO) listed within Table 3.

According to Table 8, 12 ORFs of the 416 predicted 15 coding regions were identical to public database entries (98% to 100% homology at the amino acid level), namely: y4hI (*nodA*), y4hH (*nodB*), y4hG (*nodC*), y4aL (*nodD1*), y4nC (*nodS*), y4nB (*nodU*), y4sM (ORF1), y4vK (*nifH1*), y4xA (*nifH2*), y4yD (*noIW*), y4yE (*noIB*), y4yG (*noIU*). In addition, the database 20 entry of the homologue to y4yC (*noIX*) has been corrected to 98% identical to y4yC. Furthermore, the sequence of the ORF y4hB (*noeE*) has been available to the public since October 1996. Except the 14 ORFs mentioned above, the remaining 402 ORFs are new. 139 of them show no homology to any known 25 ORF/protein. The others exhibit less than 98% amino acid identity to public database entries over their whole length.

T02280-4955E660

INDUSTRIAL APPLICABILITY

The present invention provides a detailed analysis of the symbiotic plasmid pNGR234a of *Rhizobium* sp. NGR234. The
5 plasmid pNGR234a (including any ORFs encoded therein, or any part of the nucleotide sequence of the plasmid, or any proteins expressible from any of said ORFs or any part of said nucleotide sequence) has industrial applicability which can include its use in, *inter alia*, the following areas:

10

(a) the analysis of the structure, organisation or dynamics of other genomes;

15

(b) the screening, subcloning, or amplification by PCR of nucleotide sequences;

(c) gene trapping;

20

(d) the identification and classification of organisms and their genetic information;

(e) the identification and characterisation of nucleotide sequences, amino acid sequences or proteins;

25

(f) the transportation of compounds to and from an organism which is host to at least to one of said nucleotide sequences, ORFs or proteins;

30

(g) the degradation and/or metabolism of organic, inorganic, natural or xenobiotic substances in a host organism;

35

(h) the modification of the host-range, nitrogen fixation abilities, fitness or competitiveness of organisms;

(i) obtaining a synthetic minimal set of ORFs

T02280-4966660

required for functional *Rhizobium*-legume symbiosis;

(j) the modification of the host-range of rhizobia;

(k) the augmentation of the fitness or competitiveness of *Rhizobium* sp. NGR234 in the soil and its nodulation efficiency on host plants;

(l) the introduction of desired phenotype(s) into host plants using said plasmid as a stable shuttle system for foreign DNA encoding said desired phenotype(s); or

(m) the direct transfer of said plasmid into rhizobia or other microorganisms without using other vectors for mobilization.

0939964-032701
T0230-4969660

REFERENCES

- Altschul, S.F., G. Warren, W. Miller, E.M. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
- Appelbaum, E. R., D.V. Thompson, K. Idler and N. Chartrain. 1988. *Rhizobium japonicum* USDA1 191 has two *nodD* genes that differ in primary structure and function. *J. Bacteriol.* 170: 12-20.
- Badenoch-Jones, J., T.A. Holton, C.M. Morrison, K.F. Scott and J. Shine. 1989. Structural and functional analysis of nitrogenase genes from the broad host-range *Rhizobium* strain ANU240. *Gene* 77: 141-153.
- Bender, G.L., M. Nayudu, K.K.L. Strange and B.G. Rolfe. 1988. The *nodD1* gene from *Rhizobium* strain NGR234 is a key determinant in the extension of host-range to the non-legume *Parasponia*. *Mol. Plant-Microbe Interact.* 1: 259.
- Bodmer, W.F. 1994. The Human Genome Project. *Rev. Invest. Clin. (Suppl.)* 3-5.
- Broughton, W.J., M.J. Dilworth, and I.K. Passmore. 1972. Base ratio determination using unpurified DNA. *Anal. Biochem.* 46: 164-172.
- Broughton, W.J., N. Heycke, H. Meyer z.A., and C.E. Pankhurst. 1984. Plasmid-linked *nif* and "*nod*" genes in fast growing rhizobia that nodulate *Glycine max*, *Psophocarpus tetragonolobus*, and *Vigna unguiculata*. *Proc. Natl. Acad. Sci. USA.* 81: 3093-3097.

- 35

- Fellay, R., P. Rochepeau, B. Relić, and W.J. Broughton. 1995. Signals to and emanating from *Rhizobium* largely control symbiotic specificity. In *Pathogenesis and host specificity in plant diseases. Histopathological, biochemical, genetic, and molecular bases* (ed. U.S. Singh, R.P. Singh, and K. Kohmoto), Vol. I, pp. 199-220. Pergamon/Elsevier Science Ltd., Oxford, U.K.
- 10 Fickett, J.W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10: 5303-5318.
- Fischer, H.-M. 1994. Genetic regulation of nitrogen fixation in *Rhizobia*. *Microbiol. Rev.* 58: 352-386.
- 15 Fisher, R.F. and S.R. Long. 1993. Interactions of NodD at the *nod* box: NodD binds to two distinct sites on the same face of the helix and induces a bend in the DNA. *J. Mol. Biol.* 233: 336-348.
- 20 Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.
- 25 Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.
- 30 Freiberg, C., X. Perret, W.J. Broughton and A. Rosenthal. 1996. Sequencing the 500-kb GC-rich symbiotic replicon of *Rhizobium* sp. NGR234 using dye terminators and a thermostable sequenase: A beginning. *Genome Research*, in
- 35 press.

0593964-032701
T02280-4965550

Gribskov, M., J. Devereux, and R.R. Burgess. 1984. The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* **12**: 539-549.

5

Hanahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166**: 557-580.

Hartl, D.L. and M.J. Palazzolo. 1993. *Drosophila* as a model organism in genome analysis. In *Genome research in molecular medicine and virology* (ed. K.W. Adolf), pp. 115-129. Academic Press, Orlando, FL, U.S.A.

Hiles, I.D., M.P. Gallagher, D.J. Jamieson, and C.F. Higgins, 1987. Molecular characterization of the oligopeptide permease of *Salmonella typhimurium*. *J. Mol. Biol.* **195**: 125-142.

Iismaa, S.E., P.M. Ealing, K.F. Scott, and J.M. Watson. 1989. Molecular linkage of the *nif/fix* and *nod* gene regions in *Rhizobium leguminosarum* biovar *trifolii*. *Mol. Microbiol.* **3**: 1753-1764.

Levy, J. 1994. Sequencing the yeast genome: An international achievement. *Yeast* **10**: 1689-1706.

Lewin, A., E. Cervantes, C.-H. Wong and W.J. Broughton. 1990. *nodSU*, two new *nod* genes of the broad host range *Rhizobium* strain NGR234 encode host-specific nodulation of the tropical tree *Leucaena leucocephala*. *Mol. Plant Microbe Interact.* **3**: 317-326.

Long, S.R. 1989. *Rhizobium*-legume nodulation: life together in the underground. *Cell* **56**: 203-214.

35

T02280-4966E660

- Long, S., J.W. Reed, J. Himawan and G.C. Walker. 1988. Genetic analysis of a cluster of genes required for synthesis of the calcofluor-binding exopolysaccharide of *Rhizobium meliloti*. *J. Bacteriol.* 170: 4239-4248.
- 5 Makino, S.-I., I. Uchida, N. Terakado, C. Sasakawa, and M. Yoshikawa. 1989. Molecular characterization and protein analysis of the cap region, which is essential for encapsulation in *Bacillus anthracis*. *J. Bacteriol* 171: 722-
10 730.
- Martinez, E., D. Romero, and R. Palacios. 1990. The *Rhizobium* genome. *Crit. Rev. Plant Sci.* 9: 59-93.
- 15 Morett, E. and M. Buck. 1988. NifA-dependent in vivo protection demonstrates that the upstream activator sequence of *nif* promoters is a protein binding site. *Proc. Natl. Acad. Sci. USA.* 85: 9401-9405.
- 20 Morett, E. and M. Buck. 1989. In vivo studies on the interaction of RNA polymerase- σ^{54} with the *Klebsiella pneumoniae* and *Rhizobium meliloti nifH* promoters: The role of *nifA* in the formation of an open promoter complex. *J. Mol. Biol.* 210: 65-77.
- 25 Padmanabhan, S., R.-D. Hirtz, and W.J. Broughton. 1990. Rhizobia in tropical legumes: Cultural characteristics of *Bradyrhizobium* and *Rhizobium* sp. *Soil Biol. Biochem.* 22: 23-28.
- 30 Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 85: 2444-2448.
- 35 Perego, M., C.F. Higgins, S.R. Pearce, M.P. Gallagher, and J.A. Hoch. 1991. The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. *Mol. Microbiol.* 5: 173-185.

- Perret, X., W.J. Broughton, and S. Brenner. 1991. Canonical ordered cosmid library of the symbiotic plasmid of *Rhizobium* species NGR234. *Proc. Natl. Acad. Sci. USA.* **88**: 1923-1927.
- 5 Perret, X., R. Fellay, A.J. Bjourson, J.E. Cooper, S. Brenner, and W.J. Broughton. 1994. Subtraction hybridization and shotgun sequencing: A new approach to identify symbiotic loci. *Nuclei Acids Res.* **22**: 1335-1341.
- 10 Platt, T. 1986. Transcription termination and regulation of gene expression. *Annu. Rev. Biochem.* **55**: 339-372.
- 15 Radloff, R., W. Bauer, and J. Vinograd. 1967. A dye-buoyant-density method for the detection and isolation of closed circular duplex DNA: The closed circular DNA in HELA cells. *Proc. Natl. Acad. Sci. USA.* **57**: 1514-1521.
- 20 Relić, B., X. Perret, M.T. Estrada-García, J. Kopcinska, W. Golinowski, H.B. Krishnan, S.G. Pueppke and W.J. Broughton. 1994. Nod factors of *Rhizobium* are a key to the legume door. *Mol. Microbiol.* **13**: 171-178.
- 25 Rosenthal, A. and D.S. Charnock-Jones. 1993. Linear amplification sequencing with dye terminators. *Methods Mol. Biol.* **23**: 281-296.
- 30 Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, U.S.A.
- 35 Shine, J. and L. Dalgarno. 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementary to nonsense triplets and ribosome binding sites. *Proc Natl. Acad. Sci.* **71**: 1342-1346.

003964-082701
 102280-4966660

- Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**: 195-197.
- Stanfield, S., L. Ielpi, D. O'Brochta, D.R. Hesinki and G.S. Ditta. 1988. The *ndvA* gene product of *Rhizobium meliloti* is required for Beta(1-2)glucan production and has homology to the ATP binding export protein HlyB. *J. Bacteriol.* **170**: 3523-3530.
- 10 Sulston, J, Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, et al. 1992. The *C. elegans* genome sequencing project: A beginning. *Nature* **356**: 37-41.
- 15 Tabor, S. and C.C. Richardson. 1995. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci.* **92**: 6339-6343.
- 20 van Rhijn, P. and J. Vanderleyden. 1995. The *Rhizobium*-plant symbiosis. *Microbiol. Rev.* **59**: 124-142.
- van Slooten, J.C., T.V. Bhuvanasvari, S. Bardin, and J. Stanley. 1992. Two C_4 -dicarboxylate transport systems in *Rhizobium* sp. NGR234: Rhizobial dicarboxylate transport is essential for nitrogen fixation in tropical legume symbioses. *Mol. Plant Microbe Interact.* **5**: 179-186.
- 25 Yanisch-Perron, C., J. Ira, and J. Messing. 1985. Improved M13 phage cloning vectors and host strains: Nucleotide sequences of M13mp18 and pUC19 vectors. *Gene* **33**: 103-119.
- 30

0993994-082701
102280-4965660

Bairoch A., P. Bucher, and K. Hofmann. 1995. The prosite database, its status in 1995. *Nucleic Acids Res.*, **24** 189.

Borodovsky, M.Y., K.E. Rudd and E.V. Koonin. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome *Nucleic Acids Res.* **22**: 4756.

Broughton, W.J., U. Samrey, and J. Stanley. 1987. Ecological genetics of *Rhizobium meliloti*: symbiotic plasmid transfer in the *Medicago sativa* rhizosphere *FEMS Microbiol. Lett.* **40**: 251.

Fellay, R., X. Perret, V. Viprey, W.J. Broughton, and S. Brenner. 1995a. Organization of host-inducible transcripts on the symbiotic plasmid of *Rhizobium* sp. NGR234 *Mol. Microbiol.* **16**: 657.

Freiberg, C., R. Fellay, A. Bairoch, W.J. Broughton, A. Rosenthal, and X. Perret. 1997. Molecular basis of symbiosis between *Rhizobium* and legumes. *Nature*, **387**: 394-401.

Gray, J.X., M.A. Djordjevic, and B.G. Rolfe. 1990. Two genes that regulate exopolysaccharide production in *Rhizobium* sp. strain NGR234: DNA sequences and resultant phenotypes *J. Bacteriol.* **172**: 195.

Hanin, M., S. Jabbouri, D. Quesada-Vincens, C. Freiberg, X. Perret, J.-C. Promé, W.J. Broughton, and R. Fellay. 1996. Sulphatation of *Rhizobium* sp. NGR234 Nod factors is dependent on *noeE*, a new host-specificity gene *Mol. Microbiol.*, in press.

Krishnan, H.B., C.-I. Kuo, and S.G. Pueppke. 1995. Elaboration of flavonoid-induced proteins by the nitrogen-fixing soybean symbiont *Rhizobium fredii* is regulated by both *nodD1* and *nodD2*, and is dependent on the cultivar-specificity locus, *noIXWBTUV* *Microbiology.* **141**: 2245.

Morrison, N.A., C.Y. Hau, M.J. Trinick, J. Shine and B.G. Rolfe. 1983. Heat curing of a sym plasmid in a fast-growing *Rhizobium* sp. that is able to nodulate legumes and the nonlegume *Parasponia* sp. *J. Bacteriol.* **153**: 427.

Nakai, K. and M. Kanehisa. 1992. Expert system for predicting protein localization sites in Gram-negative bacteria. *PROTEINS: Structure, Functions, and Genetics* **11**: 95-110.

- Piper, K.R., S. Beck von Bodman, and S.K. Farrand. 1993. Conjugation factor of *Agrobacterium tumefaciens* regulates Ti plasmid transfer by autoinduction *Nature* **362**: 448.
- Sullivan, J.T., H.N. Patrick, W.L. Lowther, D.B. Scott, and C.W. Ronson. 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment *Proc. Natl. Acad. Sci.*, **92**: 8985.
- van Slooten, J.C., E. Cervantes, W.J. Broughton, C.-H. Wong, and J. Stanley. 1990. Sequence and analysis of the *rpoN* sigma factor gene of *Rhizobium* sp. strain NGR234 *J. Bacteriol.* **172**: 5563.
- van Slooten, J.C., T.V. Bhuvanaswari, S. Bardin, and J. Stanley. 1992. Two C4-dicarboxylate transport systems in *Rhizobium* sp. NGR234: rhizobial dicarboxylate transport is essential for nitrogen fixation in tropical legume symbioses *Mol. Plant-Microbe Interact.* **5**: 179.
- Zhang, L.-H., P.J. Murphy, A. Kerr, and M.E. Tate. 1993. *Agrobacterium* conjugation and gene regulation by N-acyl-L-homoserine lactones *Nature* **362**: 446.

0999964-083701
FO/230-49663660